The American Economic Review

Cwk-Ho2181-88-p003534
391 83
.001

ARTICLES

W. A. LEWIS **The State of Development Theory**

R. EISNER AND P. J. PIEPER
A New View of the Federal Debt and Budget Deficits

M. D. COHEN AND R. AXELROD
Coping with Complexity

A. J. FIELD **Asset Exchanges and the Transactions Demand for Money, 1919–29**

D. FRIEDMAN **On the Efficiency of Experimental Double Auction Markets**

J. M. MARSHALL **Gambles and the Shadow Price of Death**

G. D. LIBECAP AND S. N. WIGGINS
Contractual Responses to the Common Pool

F. DE LEEUW AND M. J. MCKELVEY
Price Expectations of Business Firms

O. ASHENFELTER AND D. E. BLOOM
Models of Arbitrator Behavior

R. K. SAH AND J. E. STIGLITZ
The Economics of Price Scissors

S. HOLLANDER **Marx and Malthusianism**

C. R. HULTEN AND R. M. SCHWAB
Regional Productivity Growth in U.S. Manufacturing: 1951–78

E. R. MOREY **Confuser Surplus**

SHORTER PAPERS: D. N. McCloskey and J. Nash; G. P. Dwyer, Jr. and C. M. Lindsay; G. Fallis and L. B. Smith; W. Krämer and K. Neusser; D. E. Wildasin; J. E. Kwoka, Jr.; H. N. Goldstein and S. E. Haynes; P. Davis and G. F. Papanek; N. H. Leff; R. Ram; R. J. Gilbert and D. M. G. Newbery; J. Reinganum; S. W. Salant; R. J. Gilbert and D. M. G. Newbery; O. J. Evans; B. B. White.

MARCH 1984

THE AMERICAN ECONOMIC ASSOCIATION

•Published at George Banta Co., Inc. Menasha, Wisconsin. The publication number is ISSN 0002-8282.

•*THE AMERICAN ECONOMIC REVIEW* including four quarterly numbers, the *Proceedings* of the annual meetings, the Directory, and Supplements, is published by the American Economic Association and is sent to all members five times a year: March; May; June; September; December.

Dues for 1984, which include a subscription to both the *American Economic Review* and the *Journal of Economic Literature*, are as follows:

$33.00 for regular members with rank of assistant professor or lower, or with annual income of $15,840, or less;

$39.60 for regular members with rank of associate professor, or with annual income of $15,840 to $26,400;

$46.20 for regular members with rank of full professor, or with annual income above $26,400;

$16.50 for junior members (registered students). Certification must be submitted yearly.

Institutional subscriptions (libraries, institutions, or firms) are $100.00 a year. Individual subscriptions are $65.00 a year. Only subscriptions to both publications will be accepted. Single copies of either journal may be purchased from the Secretary's office.

In countries other than the United States, add $9.20 to cover extra postage.

•Correspondence relating to the Directory, advertising, permission to quote, business matters, subscriptions, membership and changes of address should be sent to the Secretary, C. Elton Hinshaw, 1313 21st Avenue So., Suite 809, Nashville, TN 37212-2786. Change of address must reach the Secretary at least six (6) weeks prior to the month of publication. The Association's publications are mailed second class.

•Second-class postage paid at Nashville, Tennessee and at additional mailing offices, Printed in U.S.A.

•Postmaster: Send address changes to *American Economic Review*, 1313 21st Avenue So., Suite 809, Nashville, TN 37212-2786.

Founded in 1885

Officers

President

CHARLES L. SCHULTZE

The Brookings Institution

President-Elect

CHARLES P. KINDLEBERGER

Massachusetts Institute of Technology

Vice-Presidents

ZVI GRILICHES

Harvard University

ROBERT L. HEILBRONER

New School for Social Research

Secretary

C. ELTON HINSHAW

Vanderbilt University

Treasurer

RENDIGS FELS

Vanderbilt University

Managing Editor of The American Economic Review

ROBERT W. CLOWER

University of California-Los Angeles

Managing Editor of The Journal of Economic Literature

MOSES ABRAMOVITZ

Stanford University

P3534

Executive Committee

Elected Members of the Executive Committee

ANN F. FRIEDLAENDER

Massachusetts Institute of Technology

JOSEPH E. STIGLITZ

Princeton University

WILLIAM D. NORDHAUS

Yale University

A. MICHAEL SPENCE

Harvard University

VICTOR R. FUCHS

Stanford University

JANET L. NORWOOD

Bureau of Labor Statistics

EX OFFICIO Members

GARDNER ACKLEY

The University of Michigan

W. ARTHUR LEWIS

Princeton University

THE AMERICAN ECONOMIC REVIEW

ROBERT W. CLOWER
Managing Editor

JOHN G. RILEY
Associate Editor

WILMA ST. JOHN
Production Editor

THERESA DE MARIA
Assistant Editor

LOIS J. BAGLEY
Editorial Assistant

Board of Editors

GEORGE A. AKERLOF
G. O. BIERWAG
THOMAS F. COOLEY
PATRICIA DANZON
RONALD G. EHRENBERG
H. E. FRECH III
JACK HIRSHLEIFER
LAURENCE J. KOTLIKOFF
FREDERIC S. MISHKIN
SHERWIN ROSEN
RICHARD SCHMALENSEE
E. ROY WEINTRAUB

•Manuscripts and editorial correspondence relating to the regular quarterly issues of this *Review* and the *Papers and Proceedings* should be addressed to Robert W. Clower, Managing Editor, *AER* Editorial Office, University of California, Los Angeles, CA 90024. Manuscripts should be submitted in triplicate and in acceptable form, and should be no longer than 50 pages of double-spaced typescript. A submission fee must accompany each manuscript: $25 for members; $50 for nonmembers. *Style Instructions* for guidance in preparing manuscripts will be provided upon request to the editor.

•No responsibility for the views expressed by authors in this *Review* is assumed by the editors or the publishers, The American Economic Association.

•Copyright © American Economic Association 1984. All rights reserved.

391

March 1984 ,001

VOLUME 74, NUMBER 1

Articles

The State of Development Theory *W. Arthur Lewis* 1

A New View of the Federal Debt and Budget Deficits *Robert Eisner and Paul J. Pieper* 11

Coping with Complexity: The Adaptive Value of Changing Utility *Michael D. Cohen and Robert Axelrod* 30

Asset Exchanges and the Transactions Demand for Money, 1919–29 *Alexander J. Field* 43

On the Efficiency of Experimental Double Auction Markets *Daniel Friedman* 60

Gambles and the Shadow Price of Death *John M. Marshall* 73

Contractual Responses to the Common Pool: Prorationing of Crude Oil Production *Gary D. Libecap and Steven N. Wiggins* 87

Price Expectations of Business Firms: Bias in the Short and Long Run *Frank de Leeuw and Michael J. McKelvey* 99

Models of Arbitrator Behavior: Theory and Evidence *Orley Ashenfelter and David E. Bloom* 111

The Economics of Price Scissors *Raaj Kumar Sah and Joseph E. Stiglitz* 125

Marx and Malthusianism: Marx's Secular Path of Wages *Samuel Hollander* 139

Regional Productivity Growth in U.S. Manufacturing: 1951–78 *Charles R. Hulten and Robert M. Schwab* 152

Confuser Surplus *Edward R. Morey* 163

Shorter Papers

Corn at Interest: The Extent and Cost of Grain Storage in Medieval England *Donald N. McCloskey and John Nash* 174
Robert Giffen and the Irish Potato *Gerald P. Dwyer, Jr. and Cotton M. Lindsay* 188
Uncontrolled Prices in a Controlled Market: The Case of Rent Controls *George Fallis and Lawrence B. Smith* 193
The Emergence of Countercyclical U.S. Fertility: Note *Walter Krämer and Klaus Neusser* 201
The *q* Theory of Investment with Many Capital Goods *David E. Wildasin* 203
Advertising and the Price and Quality of Optometric Services *John E. Kwoka, Jr.* 211
A Critical Appraisal of McKinnon's World Money Supply Hypothesis *Henry N. Goldstein and Stephen E. Haynes* 217
Faculty Ratings of Major Economics Departments by Citations *Paul Davis and Gustav F. Papanek* 225
Dependency Rates and Savings:
  Another Look *Nathaniel H. Leff* 231
  Reply *Rati Ram* 234
Uncertain Innovation and the Persistence of Monopoly:
  Comment *Richard J. Gilbert and David M. G. Newbery* 238
  Reply *Jennifer F. Reinganum* 243
Preemptive Patenting and the Persistence of Monopoly:
  Comment *Stephen W. Salant* 247
  Reply *Richard J. Gilbert and David M. G. Newbery* 251
Empirical Tests of the Life Cycle Hypothesis:
  Comment *Owen J. Evans* 254
  Reply *Betsy Buttrill White* 258

Errata: Declining Industries and Political-Support Protectionist Motives 260

Notes 261

Number 85 of a series of photographs of past presidents of the Association

W Arthur Lewis

The State of Development Theory[†]

By W. ARTHUR LEWIS*

Development Economics deals with the structure and behavior of economies where output per head is less than 1980 US $2,000. Whether this dividing line is somewhat too high or too low is of no significance in this context. A search for precision would send us clambering among Kravis dollars—or cause us to abandon dollars in favor of other indexes; like my favorite choice, the proportion of the labor force required in agriculture to feed the whole population; or more nebulous measurements of entrepreneurial potential. Nothing turns on these distinctions, so we need not pursue them.

However defined, Development Economics is said to be now in the doldrums, after a couple of spirited decades. It seems to be true that the subject has been deserted by American Ph.D. students. Their antennae tell them where to find the best jobs. In this contest, Development Economics no longer competes. Foreign aid has been cut, the multilateral institutions cannot keep up with the inflation, and the Ford Foundation has changed its priorities. At the graduate level, our subject still beckons only one group. The number of Third World students in the United States still runs high, and they are thirsty for our subject; so universities that admit a lot of Third World students continue to have flourishing teams of development economists.

This counting of heads is liable to lead us astray. There are more students and teachers of Development Economics than are revealed by counting those who are denominated as such. To some extent, the practitioners of other disciplines have taken over from the Development Economics team those parts of the subject that run with their disciplines. Thus dual labor markets are standard in the Labor Economics Department; the International Finance Department handles the effects of disequilibrium in small, open and poor economies; and the Industry Department is studying multinational corporations and appropriate technology. This has both advantages and disadvantages. Monetary disequilibrium in poor countries should be studied both against the background of disequilibrium in rich countries and also against the general background of poverty. There is therefore no substitute for the development team that concentrates on poor countries as such, though this may not be a large team if other departments are committed to including the problems of poor countries in their teaching and research programs.

However, the reference to the doldrums goes beyond counting heads to the subject matter itself, which is said to be producing fewer new ideas (right or wrong) than it was doing in the 1950's and 1960's. The current output of factual analysis is probably larger than ever, considering the number of English language journals that it keeps going. It is the current output of new development theories that is smaller. The 1950's and 1960's were a period of great theoretical innovation and controversy.

This indictment is probably true, though not particularly damning. In any field of thought the output of worthwhile innovations is subject to fluctuations. Quiet periods are needed for digesting the great feasts. A scientific field should not be abandoned because it has not had a brilliant new idea for a decade; for all we know, it may just be about to erupt.

The principal question I want to tackle in this paper is whether there is need for a separate economics—Development Economics—for countries with less than 1980 US $2,000 per head: never mind where the students are to come from, nor how the subject

---

*Woodrow Wilson School of Public and International Affairs, Princeton University, Princeton, NJ 08544.

[†] Presidential address delivered at the ninety-sixth meeting of the American Economic Association, December 29, 1983, San Francisco, California. I have benefited from conversations with Mark Gersovitz.

matter is cross classified. This adds up to asking whether such countries differ in structure or in behavior from the richer countries, in ways that require different concepts or tools to understand their functioning.

We are not going to find an unbridgeable gap between Development Economics and the Economics of the Developed, any more than a pediatrician would claim that geriatrics was an unrelated body of knowledge. The overlap between the developing and the developed is bound to be great. The more so in economics because the two basic tools, Supply and Demand and the Quantity Theory of Money, will take you a very long way if you just want to understand what is going on. This is why there are so many good untrained economists, and also why some of our most high-powered colleagues perform no better in practical matters than a good undergraduate. Given the power of these elementary tools, one could study either the economies of the developed as preparation for studying the developing, or the economies of the developing as preparation for studying the developed. However, the differences are also rather large, which is why each also has some tools of its own.

I shall divide our subject matter into two parts: matters relating to the allocation of resources in the short run, and matters relating to long-term growth.

I. Allocation in the Short Run

In this department the differences are of degree rather than of kind; that is, developing and developed are both affected, but developing more so.

The first and easiest category for comparisons consists of cases where price does not equate to real social cost. This is easy because the problem is much the same in rich and poor countries, and is tackled with the same benefit-cost apparatus in both cases. Leading causes of price defects include externalities, Manoilesco wage differentials, learning factors, inappropriate exchange rates, inappropriate interest rates, unemployment and disguised unemployment. Economists argue about particular cases, whether the model is valid, and even if valid, whether it accords with the facts, but the legitimacy of the category is not in question. These are the building blocks of the activity known as development planning, which can be bought at any level of complexity from the benefit-cost of, say, a few miles of road to the highly esoteric general equilibrium models that are the cutting edge of the neoclassical system.

One should, however, note that the neoclassical system of price formation is under attack, even when its prices are Pareto optimal, on the ground that they may be inequitable. Economists are used to shrugging this off when making comparisons inside a given country, but the same argument has now surfaced in the analysis of trade between rich and poor countries, which yields unacceptable terms of trade because tropical produce reaches the market at prices that reflect alternative cost rather than equal pay for equal work. A whole chain of demands for a new international economic order hangs from this pillar.

My second category is again a regular part of neoclassical economics, where the unregulated market constrains productive capacity.

Location theory yields an important example, namely the formation of growth poles, where a useful tendency to form towns runs away into the creation of huge concentrations, holding many millions of people, far beyond economies of scale into the diseconomies of congestion. This is due to externalities, to prices falling under monopolistic control, and to inappropriate government actions which prevent the costs of congestion from turning up in marginal costs. The result is destruction of growth potential elsewhere; smaller production for the same total of resources than one would have if the resources were distributed over more growth centers.

The dependency concept in international trade theory is related to the growth pole. In either case, potential is destroyed at the periphery. The market works to concentrate rather than to diffuse the benefits of trade.

One may mention here, as another example of market failure, the tendency of small farmers to become heavily indebted; whereupon their incentives to invest in raising the productivity of their lands is reduced. The well-known remedies, such as prohibiting the mortgaging of agricultural land, substitute one problem for another.

I have been talking about market failure to move towards the right equilibrium. The failure to move smoothly even when right is another distinction between rich and poor countries. This derives from low elasticities of demand and supply, coupled with low levels of stocks of commodities in the poor countries; as a result of which market events bring quick and large changes in price, and only slow responses in quantity: the phenomenon known as "structuralism." These wide price ranges are disruptive in production and also in the social tensions exacerbated by resulting disturbances in income distribution. Governments are then driven, when dealing with commodities of mass interest, to use quotas, rations, licenses and other market-sharing arrangements.

What I have said so far about the functioning of the price mechanism does not distinguish rich and poor countries in theory, though it does so in practice. For one thing, if these points are listed as weaknesses of market systems, they are more prevalent in poor than in rich countries. Wherefore, secondly, practitioners of the Economics of the Developed get into the habit of writing as if the prices yielded by market systems were obviously the right prices, and all other prices are "distortions"; whereas development economists get the opposite habit of writing as though nothing mattered but these blemishes. Thirdly, the theoretical work has been done by development economists, even when the new model is used for analysis of developed economies. The list of new models invented by development economists in the 1950's and 1960's is quite impressive:

two-gap model
unbalanced growth
vent for surplus
Dutch disease
dual economy
disguised unemployment
structural inflation
dependency
indicative planning
appropriate technology
big push
growth pole
rising savings ratio
low-level equilibrium trap

It is unlikely that students will get to study this set of models unless separate provision is made for teaching Development Economics.

There remain *two categories* of cases where the assumptions of market economics may not hold. The first is almost exclusively a phenomenon of poor countries. Here some production and exchange are governed not by desire for income maximization, but by other "noneconomic" considerations. Thus in some countries living near subsistence level, production and exchange are governed by ritual laws, based on kinship and on authority status. Anthropologists love those islands where the most successful fishermen burn some of their boats to demonstrate their superior achievement. I draw attention not to acts of conspicuous consumption, which are universal, but to deliberate destruction of productive assests, which is not.

Members of a community may also be divided by language, race, religion, or other tribal division. When a utility company plans to build a dam in Province *A* to feed electricity to a town in Province *B*, the cost and the flow of benefits in the building of the dam will depend on the technology applied. The dam may cost more in money when built by unskilled labor than when built with imported bulldozers, yet the benefit-cost analysis may show the cost with unskilled labor cheaper, whether because of divergence between money wages and real labor cost, or because current exchange rates or interest rates are inappropriate. In the economics of rich countries, we do not distinguish between men of *A* and men of *B*, but, in plural societies, *A* cannot be added to *B* by simple arithmetic. The benefit-cost analysis may have to list gains and losses separately by ethnic classification. Perhaps a plural society will use bulldozers, where a homogeneous society would use unskilled labor.

We rely on anthropologists for our knowledge of such matters, and economic anthropology is a speciality on its own. It has always attracted more anthropologists than economists, and links have diminished since economics moved beyond supply and demand and the Quantity Theory. Development economists are no more likely than other economists to move into economic anthropology.

Of course, the argument for an anthropological approach to economic behavior is not confined to poor countries. The study of economic behavior in rich countries would benefit from more observation (statistical as well as anthropological) rather than more deductive reasoning. Large elements of human behavior—investing, saving, having children, bearing or avoiding risks, setting attainable targets—are not dominated by the calculus of marginal utility, unless utility is given a universalistic and therefore useless meaning. Development Economics, which relies more on history, statistics, and anthropology, may to this extent be a little more realistic than other economics.

The second noneconomic factor that development economists cannot leave out of their calculations is the government's behavior. Although government expenditure is a smaller share of national income in poor countries, nevertheless the government is more closely intertwined in poor countries with the modern sector, where no large private event occurs without the government becoming involved in some way. Hence an economic analyst who tries to figure out the likely consequence of any economic event must assess how the government may react. We usually say that such situations require an interdisciplinary approach. Since in practice there is no interdisciplinary discipline, and since the economist's counsel is sought separately from that of political experts, the economist is forced *faute de mieux* to become his own political scientist. It then soon emerges that we need a sound political science as much as a sound economic science.

One reason why government is closer to the developing economy is that the market works less efficiently there than in the developed economy. So government is constantly asked to rectify market error or market inequity. To say this is not to imply that government action in the market always gives a better answer than the uncontrolled market, whether in allocation or in distribution. For just as the market gives better service in rich than in poor countries, so also the government tends to be better administered, has more resources, and is slightly less dominated by "politics" and personal advantage in decision making. Pigou taught economists that the market was imperfect, but that points of imperfection could usually be eliminated by handing the problem over to the government for tax or subsidy. Such an assumption is not valid for all *LDC*s. Here it is often the case that the imperfect solution of the market could be better than that of the government. The government needs to be modernized just as much as the market. Also the assumption that the government "represents" the people may not hold. Political scientists offer us many different models of government-military (with generals), military (with sergeants), technocratic, aristocratic, popular front, peasant, kleptocratic—which react differently to similar stimuli.

To sum up, in matters relating to the allocation of resources in the short run, Development Economics marches most of the way with the economics of developed market economies, except that there are more special cases; and except that sociological implications cannot be brushed aside, or be assumed to be held constant while decisions are reached on the basis of "economic" considerations alone.

II. Long-Run Growth

The matters at issue fall into two categories: the Search for the Engine of Growth and the Patterns of Growth.

The economist's dream would be to have a single theory of growth that took an economy from the lowest level of say $100 per capita, past the dividing line of $2,000, up to the level of Western Europe and beyond. Or to have, since processes may differ at different stages, a set of theories growing out of each other longitudinally, and handing over to each other. Or putting aside what happens after $2,000 is passed, to have at least one good theory for the developing economy from $100 to the dividing line.

We are well endowed with models for the final state of economic maturity. Most schools of economics have projected their models into the future. The results are usually discouraging. The model has some constraint which fails to expand in proportion with the rest of the resources, and therefore pulls the growth rate down. Adam Smith's economy grows to its "full complement of riches," as permitted by its laws and institutions, and there it rests. Ricardo's model is brought to stagnation by the exorbitant rise in rents. Marx's model is floored by continuously increasing technological unemployment, due to inadquate investment. Cassell's neoclassical model comes to stagnation because capital accumulation drives interest rates too low. Later neoclassical models escape retrogression. By eliminating land and maneuvering the growth of capital to equality with the growth of population, one can then obtain steady states with or without the Golden Rule.

Since we are not interested in the fate of mature economies, we can pass by this set of theories, and begin instead with the lowest levels of income.

We are also well endowed with theories for this end of the scale. The first problem is why so many countries have grown so slowly over long periods, and especially since 1850, when the main elements of modernization were already well understood. The second problem is why, after a while, countries do join the mainstream and float up past $2,000 per capita. We must be able to explain both what inhibits and what releases energy.

One can begin with threshold models, which require a country to attain some minimum level before it can move towards continual growth. Rostow's model, which is also the last of the "stages of growth" models, has this period in which the "pre-conditions for take-off" are established. The low-level equilibrium trap is another example. The "big push" is a third.

Other obstacles in the way of movement include negative externalities, the two-gap model, and the need for balanced growth.

Then there are the models of paralysis, where entry into the world market is debilitating. Thus a country may be drained of its human and physical capital by a nearby growth pole. Or it may suffer the pangs of dependency. Or its potential labor supply may be reduced by inhuman treatment by colonial employers.

Since World War II, the developing countries growing least rapidly (the bottom third, growing at less than 3½ percent) have had one or other of two handicaps. Either their rainfall is inadequate, so that their agriculture contributes little to growth. The farm surplus is too small. Senegal is an example. Or else they have had significant internal political disturbance, which has discouraged investment by their own natives, no less than by foreigners. This is what has shown up over two decades of world prosperity; whether the same differentiation would show over a period of world depression we cannot say.

No-growth models are going out of fashion. Their logical validity was never in doubt; at issue was their relevance. Two forces have worked against them. First, there has been so much growth among *LDC*s that models proving that growth could not happen are now less persuasive. Secondly, these models were anaesthetic. One could examine a country and pronounce it not to be growing without hurting the feelings of its inhabitants; one model or another could be used to show that it was not their fault. Nowadays it is accepted that things sometimes go wrong because rulers have adopted mistaken policies, and we do not need the anaesthesia.

Fashions aside, what the models set out to explain is real: the long near stagnation of so many countries. Equally real is the fact that economies escape from this stagnation, raising the question of why some move sooner or faster than others.

In its present state, our material on the middle-income countries may be divided into three categories of theories: for the growth of resources, for the patterns of growth, and for the processes of investment.

Our theories for the growth of resources in middle countries are building up. For the size of the labor force, we have population theories, labor force participation, dual markets, human capital, education, and all that. We overlap with the study of such matters in advanced economies, though our results are not always the same.

We are collecting material on capital in low-income countries—theories of savings, private and public. Here we are at odds with students of developed countries, whose objective was to explain why the savings ratio is constant, where our objective is to explain why it rises.

We are also at odds on the quantity and quality of entrepreneurship. After much disputation *MDC* economists have abandoned this subject, taking the line that a profitable economy will have no shortage of entrepreneurs. We, on the other hand, observe that the Ibos show more entrepreneurship their countrymen the Tivs, or the Chinese than their countrymen the Malays. We also observe for the Third World as a whole a level of foreign entrepreneurship which is resented despite the obvious benefits deriving from its presence.

This part of the subject—the growth of resources—is lively and needs no defense.

The second set of materials relates to the patterns of growth. We know more about the patterns imposed by growth than we know for certain about the causes of growth. This part of our subject has a long history, reaching back to Petty's observations about changes in the relative proportions of the labor force in agriculture and services. Interest was revived by Colin Clark's book on the *Conditions of Economic Progress*, since when we have had a large literature led by Simon Kuznets and Hollis Chenery.

The topic has spilled over to calculate other kinds of sectoral changes: the sectoral composition of imports and exports; the share of income and nonincome taxes in government revenue; etc. Any propensity that is thought to be a function of income per head can be derived, whether from longitudinal or from cross-country observations. The technique is particularly helpful for those engaged in development planning. It has, of course, its weaknesses, including the fact that cross-country and longitudinal studies frequently give quite different answers. So the results must be used with care.

The evolution of the urban labor market is a particularly interesting part of the pattern of growth: the emergence of noncompeting groups, the bifurcation into superior and inferior tracks, the creation of reserves of surplus labor in disguised or undisguised unemployment. All this is occurring now in front of our eyes. Here *MDC* and *LDC* economists have fertilized each other.

The pattern of growth is at the heart of the argument as to alternative strategies of growth, since different strategies yield different patterns. The approach of the economist, qua philosopher, is to try to understand what is going on; the approach of the "social engineer" is to change the outcome. At issue is the percentage of the national income received by the bottom layers of the population, perhaps the bottom 80 percent, since, in *LDC*s, the man at 20 points from the top is pretty poor. The package for the rural poor is fairly well agreed—embracing land reform, irrigation, fertilizers, better varieties and the usual agricultural extension measures, so that they may produce more, plus a social service package of schooling, public health and such, which can count either as increased consumption or as investment in human capital. The urban package is not so obvious or agreed, since one of its components, nationalized industry, has in low-income countries doubtful effects on output per worker; but this package also includes more food and social services.

Looked at from the angle of consumption, this is a "Basic Needs Strategy." It says that a wider distribution of income must be a social objective along with a larger output per head. Its claim to be an "Alternative Strategy of Development" is presumably based on the relatively greater expenditure on social services, which will also count as investment in human capital. Proposed changes in structure, such as distribution of land or emphasis on small-scale industry, are not unique to an alternative strategy.

Insofar as the philosophers are concerned, the crux of the matter is how output would be affected by policies that gave say an extra five percentage points of the national income to the bottom 80 percent of the population, assuming a peaceful transfer over, say, 10 years. Would output per head rise faster, more slowly, or at the same rate? One must also ask whether it matters, or should the change be made in any case?

The priority accorded to such questions is determined by politicians in the *LDC*s, and not by economists in the *MDC*s. Our contribution is to assess the facts and analyze them as dispassionately as we can. This we have not failed to do.

III. The Engine of Growth

The third set of materials relates to the incentives for investment, whether in physical or in human capital. It is not hereby argued that investment is the sole contributor to growth, but it is highly correlated with growth, and may serve as a proxy for the forces propelling the economy. Here we need material on the adaptation and diffusion of new techniques; on credit institutions, and such. The management of the economy is vital, that is to say the maintenance of a mutually consistent set of exchange rates, interest rates, reserve ratios, agricultural prices and wages. We need much more research in this area, so it is not surprising that many developing countries go down to defeat here. Another major cause of defeat is found where the political conditions are not consistent with new investment.

The relationship between incentives and institutions is one of the oldest parts of Development Economics. Our forefathers worried about the institutional framework of agricultural societies: about land tenure, primogeniture, sharecropping and such; and later formed quite sharp opinions about the joint stock company as it came along. The important point that differentiates them from the spirit of our age is that they did not expect to reach conclusions in this area from first principles only. Their writings about such matters are as much historical as economic analysis. And their conclusions recognize that the same question requires different answers at different times or in different places, and cannot therefore be answered without reference to circumstances.

I make these remarks, which are obvious, not for completeness but in order to point up one of the weaknesses of our subject, namely the widening gap between Economics and Economic History in Development Economics. If our subject is lowering its sights, this may be because the demise of Economic History in economics departments has brought us a generation of economists with no historical background. This is in marked contrast with the development economists of the 1950's, practically all of whom had had some historical training, and guided by Gerschenkron and Rostow, looked to history for enlightenment on the processes of development.

Given the importance of incentives and institutions, are there particular circumstances that favor growth? If economists have not failed to theorize about mature economies and about the poorest economies, neither have they failed to account for the in-betweens. Every school has offered its own candidate for driver of the engine of growth. The physiocrats, agriculture; the Mercantilists, an export surplus; the classicists, the free market; the Marxists, capital; the neoclassicists, entrepreneurship; the Fabians, government; the Stalinists, industrialization; and the Chicago school, schooling. Speaking for the geographers, one should add that it helps to have rich minerals or well-watered soils; and for the political scientists, one should add enough political security to encourage the indigenous peoples to invest in their own country.

We have it over our predecessors that we can apply econometric techniques to the search for the engine of growth. However, our answers are no more satisfactory. Application of Denison's technique of growth accounting yields the same stone for *LDC*s as for *MDC*s, namely a large residual. And Tinbergen-Klein models, whose forecasts become less reliable over more than half a year, despite the massive volume of statistics which they use in *MDC*s, are no better in *LDC*s.

We know why these income-projection models do so poorly in the long run. The reason is because success is determined not by the increase in the physical factors of production, but by a set of intangibles—political security, the quality of the infrastructure, the reliability of skilled workers and contractors, the opening up of new market opportunities, and so on. Adelman and Morris try to capture these intangibles, but run into problems of identification.

Given the range of possibilities a search for "the" engine of growth must be foredoomed. Growth occurs wherever there is a gap between capability and opportunity. Capability covers skills (domestic and foreign), government, savings, and technology. Opportunity can be of any kind, including markets, rainfall, access to licenses, infrastructure. The engine may be at home or abroad, an innovation, a good site for a transportation center, or much else.

It follows that a model for an economy at this level must be rather complex. There is no one growth theory, but a set of complementary theories. We do not yet have agreement on this set. At the core must be a theory of distribution, since this is going to provide incentives and savings. But economists cannot agree what determines the distribution of incomes in Western Europe, so how could we predict what effects growth would have on the distribution of income in Thailand? Here we should also put what we know about the growth of the volume of resources—the labor force, capital, knowledge.

Another building block would be a theory of government, where government would appear to be as much the problem as the solution. There must be a theory of class formation and class conflict, to go hand in hand with a theory of entrepreneurship. Thus a theory of the growth of the economy as a whole brings together what we know of its parts.

Such reasoning is not acceptable to all. A first line of attack is against this eclectic procedure: that this conglomeration of theories of various kinds does not make a unified subject for study. The material is unified by its subject matter—the behavior of low-income economies—and by its tools—especially supply and demand and statistical analysis. The problem is not how to take over relevant history or sociology or anthropology, but, on the contrary, how to avoid rushing in with economic answers beyond the limits within which they apply. These limits are different for *MDC*s and *LDC*s. As I said earlier, there is an element of utility maximization in studying fertility, rates of saving, migration or entrepreneurship but it tends to be overladen in poorer countries by social, religious, and political rules, that must be approached through the disciplines of political science and social psychology. The argument is not about quantification; everybody supports making historical statistics available, and the rapid increase in materials suitable for longitudinal or for cross-country statistical analysis is also a bonus for all. At issue is the working of institutions that influence decisions, and the changing of these institutions through time. Also at issue is the kind of thesis that we demand from Ph.D. students. Development Economics will certainly die if they come to think, rightly or wrongly, that work on economic institutions will not count for distinction in Ph.D. exams.

The second line of attack is that this approach abandons predictability, where the future is thought to be predictable. Not all predictability is abandoned: we can predict that the birth rate will fall, that meat consumption will rise and so on, but we cannot predict the future of the economy as a whole. What lacks is the vision of the future, already predetermined by history—or, to change the metaphor, a theory of development that is like the life story of a horse, beginning as a tiny speck and developing all its life in accordance with its genetic code already imprinted at conception. Instead we offer a theory of development that is more like the manufacture of a car, buying windshields here, tires there, carburetors some other place. The biological theory, which is "natural," sounds more attractive than the assembly line theory, which is "artificial." But cars run faster and longer than horses.

Many people have had the vision and offered us long term predictions, notably Marx, Rostow, and Toynbee. But most social scientists have rejected such models on principle. They have not discovered whether history has a purpose, or if it has, what timetable it is working to.

Looking closer in, one would particularly like to be able to predict, at a much lower level of speculation, "self-sustaining growth" (to use Rostow's phrase for development).

Many countries grow for several years at 5 percent or more, after which they grow slowly or may even decline. So it is not enough to have a model that produces growth; one must also be able to explain why some countries fall out of the line while others keep up the pace.

The analysis of self-sustaining growth proceeds at two levels; one to do with resources, and the other to do with leadership. At the resource level, a country may be judged to have reached self-sustaining growth when it is more or less self-sufficient in savings, in its managerial cadre, in skilled workers, and in other infrastructure. The physical part we can quantify, even if rather arbitrarily; say its net capital inflow is less than 2 percent of national income; private consumption is down to 60 percent of *GDP*; annual output of high school graduates is at least 5 percent of the cohort; expatriate managers and employees are less than 1 percent of the labor force; and so on. How countries get to these benchmarks is what Development Economics is about.

What cannot be predicted or prescribed is the quality of leadership that will see a country through its crises; such that if, for example, it loses an export market it will not decline, but will make for itself new economic space. This is the kind of leadership, private and public, that we expect of the countries of Northwest Europe, and it may be that most *LDC*s are too young in independence to have the cohesion that it takes. But age is not the only requirement, as Latin America testifies.

In trying to understand how a society, as distinct from its individual members, learns, chooses new directions, creates new loyalties, faces up to costly tasks, and so on, we come up to the limits of the mechanical analogies normally used by the economists, and find ourselves along with the sociologists seeking helpful biological analogies. Economists have talked about this since Alfred Marshall, but without making much impression. I doubt that philosophers are about to reach agreement on what determines the spirit of a nation, so Development Economics should stop short of that question.

Now I am ready to approach the question with which I began: do we need a separate economics for the growth of poor countries from that which we have for the growth of rich countries? The answer must be taken in parts, according to what you mean by theory of growth. If the reference is to mathematical Growth Economics, everybody agrees that this is a separate subject overlapping rather little with Development Economics. If we mean making Tinbergen models of an economy in order to predict its movement over the next twelve months, we do this in both branches, with the same less than handsome success; presumably our forecasts will all improve as we get a better understanding of the determinants of changes in investment and *MV* over the short run. The revival of trade cycle analysis will bring us closer together here. We also share a third collection of growth theory: namely the determinants of growth over 25 to 50 years, say in Britain or France or the United States for the developed, or in Ceylon or Ghana or Mexico for the developing. Here we are still lost when trying to explain why the United States sank so low in the 1930's, or what the British were doing around 1900, or why Japan leapt forward in the 1950's: with parallel questions about the wide range of performance of underdeveloped countries today, of whom one-third have been growing continually at more than 5 percent a year—an outcome without precedent. This kind of question belongs as much to history as it does to economic analysis. Development economists will owe their insights to both and will welcome any useful tools that the economists of the advanced countries may occasionally fashion for this area.

In conclusion, the concern referred to at the beginning of this paper as to the state of Development Economics seems to me quite misplaced. One's attitude must depend on one's expectations. If the critic is seeking a clear vision of whither history is leading us, he is bound to be disappointed. But if his interest is in economists collecting facts, making testable theories and testing them, then the development economists seem to be as fully engaged and as successful as any

A New View of the Federal Debt and Budget Deficits

By ROBERT EISNER AND PAUL J. PIEPER*

Much again has recently been made of the purportedly large government debt and deficits. To some, the sheer magnitude of the federal debt, passing $1.3 trillion in 1983, suggests the danger of some kind of national bankruptcy. To others, the large and growing budget deficits, which have now reached 6 percent of Gross National Product, are a harbinger of a new acceleration of inflation. Fiscal policy, it is further argued, has been incessantly stimulative, as is indicated by the movement to deficit of even the high-employment budget in the last two decades.

A number of recent papers have pointed to the need for adjustment of conventional measures of government debt and, in particular, the deficit.[1] As is of course well known, in its failure to distinguish between current and capital expenditures, the federal budget is widely at variance with prevailing accounting practice for private business. Application of government accounting methods to the private sector would move many major American corporations toward deficit. Conversely, conventional private accounting methods would generally move government accounts toward surplus.

But of further and critical importance, in government accounts as in private accounts, inflation plays vast tricks. If we are to avoid the cardinal economic sin of money illusion, we must look at real market values of debt and real deficits that correspond to changes in the real value of debt. In the first substantive section of this paper, therefore, we develop and present measures of the real value of federal government net debt, offering as a byproduct associated measures of net worth.[2] The net revaluations calculated in developing a time-series of real net federal debt are applied to develop a series of "adjusted" budget surpluses and deficits.

The second substantive section of the paper entails an application of these measures to develop adjusted figures for the recently revised official high-employment budget surplus and deficit.[3] The official and adjusted series for the high-employment budget are then related to changes in real Gross National Product and the rate of unemployment. It is found that lagged series of the official high-employment budget surplus and, even more clearly, our adjusted measures taking into account the impact of inflation and changing interest rates on the real value of the debt, show clear and sharp negative relations with the real path of the economy. With increasing inflation, a balanced official high-employment budget proved less and less consistent with the maintenance of high employment or adequate economic growth. Recent declines in real *GNP* and increases in unemployment are remarkably closely associated with a substantial move to surplus in the adjusted high-employment budgets.

---

*William R. Kenan Professor of Economics, Northwestern University, Evanston, IL 60201, and Assistant Professor of Economics, University of Illinois, Chicago, IL 60680, respectively. We are indebted to Elizabeth Fogler of the Flow of Funds Division of the Federal Reserve Board, and to John J. Seater, W. Michael Cox, and Eric Hirschhorn for guidance through published and unpublished data. We are particularly indebted to Laurence Kotlikoff for suggesting that earlier work of Eisner (1980) be updated for use by the Council of Economic Advisers. We are further indebted to the Council for financial support and to the National Science Foundation (grant SES-7717555) for support of the initial research. We are grateful to Otto Eckstein, H.S. Houthakker, and Allen Sinai for comments on an earlier draft of part of this paper, and to Moses Abramovitz, Michael Boskin, E. Cary Brown, Paul Evans, Benjamin Friedman, Robert J. Gordon, Bert G. Hickman, Frank de Leeuw, Frederic Mishkin, Dale T. Mortensen, and George Neumann for comments on later versions, and to anonymous referees. We alone are responsible for its contents.

[1] See, for example, Michael Boskin (1982), Michael Cox and Eric Hirschhorn (1983), *Economic Report of the President* (1982), Brian Horrigan and Aris Protopapadakis (1982), Rudolph Penner (1982), Richard Ruggles and Nancy Ruggles (1982), and John Seater (1981).

[2] Measures including state and local government are presented in Appendix A.

[3] See Frank de Leeuw and Thomas Holloway (1982).

TABLE 1—MARKET TO PAR INDICES (100 = PAR), SELECTED YEARS[a]

	U.S. Government Securities				Bonds		Mortgages	
	Total	Bills	Notes	Notes; Bonds	State; Local	Corporate	State; Local	Federal
1946	102.71	99.95	100.47	103.58	100.30	98.14	103.90	104.10
1959	94.07	98.80	98.00	91.73	96.00	83.77	94.50	95.60
1969	92.04	97.69	94.99	89.23	88.20	72.12	92.30	93.40
1976	101.15	98.68	103.42	102.72	101.80	97.03	96.70	97.80
1977	98.31	98.10	99.64	98.43	102.30	93.13	97.60	98.40
1978	95.30	97.15	95.80	94.38	97.40	86.43	94.50	95.90
1979	94.44	96.64	95.47	93.38	95.40	80.40	89.80	91.20
1980	92.86	96.00	94.34	91.29	86.10	74.04	86.80	88.50

[a] Complete tables corresponding to this table and to Tables 3, 4 and 5, with data for all of the years from 1946 to 1980, are available from the authors on request.

I. Actual Budget Deficits and Debt

A. *The Debt: From Par to Market and Gross to Net*

Government debt figures usually reflect the par value of government obligations. If the debt were overwhelmingly in very short-term securities, or if interest rates did not change much over the life of securities, the difference between par and market values would be small. Particularly in recent years, neither of these conditions was met. Interest rates soared and market values of outstanding fixed securities collapsed.

As a first step, therefore, we have built upon work of John Seater (1981) and Michael Cox and Eric Hirschhorn (1983) to construct market-to-par indices for major financial components of government balance sheets. We find that, with par equal to 100, the market index for the composite of all U.S. government securities had fallen by the end of 1980 to 92.86, as shown in Table 1. (The index for state and local government bonds fell further, to 86.10.)

A second step is to recognize that the gross public debt is a correct measure of neither the total liabilities (interest-bearing and non-interest-bearing) of the federal government (including its associated credit agencies and the Federal Reserve) nor of the net debt, that is, the difference between total liabilities and total financial assets. We find, as indicated in Table 2, that total liabilities associated with the U.S. government came at the end of 1980 to $1,220 billion at par. The par value of financial assets amounted to $592 billion. Par value of *net* government debt was thus only $628 billion.

But the *market* value of total U.S. government liabilities was $1,154 billion, $66 billion less than its par value. The market value of U.S. government financial assets was $707 billion, or $115 billion more than par, reflecting a market value of gold $141 billion greater than the statutory figure based on the $42 per ounce at which it is held in Treasury accounts. Thus, the market value of U.S. government net debt in 1980 was $448 billion, a considerable cry from the original "gross public debt" figure of $930 billion.[4]

---

[4] In addition to explicit debt or accounting liabilities, the federal government and its agencies of course have substantial other contractual obligations. As of the end of the 1982 fiscal year, the Treasury listed "liabilities" accountable to the public of $1,085.5 billion, "undelivered orders" (mainly of the Departments of Defense and Housing and Urban Development) of $474.7 billion and, dwarfing both, "contingency" obligations which summed to $6,981.5 billion.

These last comprised $346.7 billion in loan and credit guarantees and $2,079.6 billion in insurance commitments, largely for the Federal Deposit Insurance Corporation and the Federal Savings and Loan Investment Corporation. They also included amounts for "annuity programs," which totalled $4,473.8 billion. Of these, $835.9 billion were for retirement pay, chiefly for the armed forces and the civil service. The other major items were the actuarial values of net obligations for the three Social Security Trust Funds, coming to $577.9 billion for hospital insurance, $2,297.6 billion for old age and survivors insurance, and $656.3 billion in net *assets* for disability insurance. Adding together all of the figures

TABLE 2—THE BUDGET DEFICIT, GROSS PUBLIC DEBT, AND MARKET VALUE OF NET DEBT, 1980 (Billions of Dollars)

A. Reconcilation of Gross Public Debt and Market Value of Net Debt	
Gross Public Debt	S 930.2
+ Other Liabilities of Federal Government	120.1
+ Liabilities of Credit Agencies	188.2
+ Liabilities of Federal Reserve	173.8
− Debt held by Federal Government	− 192.46
Total U.S. Government Liabilities at Par	$1,220.0
+ Par-to-Market Conversion	− 65.7
Total U.S. Government Liabilities at Market	$1,154.3
Federal Government Financial Assets	$ 225.3
+ Credit Agency Assets	192.54
+ Federal Reserve Assets[a]	173.8
Total U.S. Government Financial Assets at Par[a]	$ 591.7
+ Par-to-Market Conversion	115.0
On Gold	141.1
On Other Assets	− 26.1
Total U.S. Government Financial Assets at Market	$ 706.7
U.S. Government Net Debt	$ 447.5
B. Reconcilation of Budget Deficit and Changes in Gross Public Debt	
Budget Deficit on National Income Account	$61.2
+ Insurance Credits to Households	+ 8.8
+ Change in U.S. Financial Assets	+ 26.0
+ Change in Treasury Debt Held by U.S. Government	+ 7.4
− Change in other U.S. Liabilities	− 11.0
− Mineral Rights Sales	− 6.5
− Statistical Discrepancy (Flow of Funds)	− 0.8
Change in Gross Public Debt	$ 85.1

[a] Includes small amount of tangible Federal Reserve assets.

Like any business, the federal government also has some tangible assets. These include reproducible assets in the form of residential structures, nonresidential structures, equipment and inventories, which by BEA estimates at the end of 1980 had a net replace-

---

(which the Treasury does not do), we come to a total of $8,541.7 billion, far in excess of direct liabilities and our net debt figures.

Such vast government programs and obligations may well have a substantial impact on the economy. They are clearly, however, in varying ways different in character and implications from the formal, current debt. A considerable amount of recent work has endeavored to detect effects on consumption and saving of the major item of implicit Social Security debt. But even measures of the magnitude of that debt vary widely with plausible assumptions as to future government policy and as to public perceptions and expectations, as has been well shown by Dean Leimer and Selig Lesnoy (1982). Eisner (1983) has also noted the apparently varied impact of "net Social Security wealth" in a macroeconomic model, along with a substantially greater and clearer role for direct government debt. As pointed out by the U.S. Treasury, "Clearly, there is a vast difference between items ... where the liability is certain and relatively precise ... and ... where the possible future liability is highly speculative and may never arise .... The [latter] amounts, if they can be projected at all, are stated ... without regard to probability of occurrence and without deduction for existing and contingent assets which would be available ... ." (1982, p. 2).

It is clear that construction of relevant measures of government commitments that go beyond net current liabilities would take us far afield. It would also be subject to huge uncertainty regarding the anticipated tax receipts which may eventually be generated to meet the commitments. One not entirely implausible assumption, particularly after recent changes in the Social Security Law, may be that we can expect taxes to be raised so that the net actuarial value of government obligations and the taxes to pay for them will be zero, and certainly not a growing ratio of national income.

In any event, the federal liabilities calculated in this paper, as in the Treasury tabulations, do not include the present value of government personnel retirement provisions, the projected Social Security commitments under current law, or other "contingent" obligations, nor do assets include an estimate of the present value of associated future taxes. Liabilities do include, however, the value of assets held in federal employee and railroad retirement funds, so that their net contribution to the debt comes to zero.

TABLE 3—FEDERAL GOVERNMENT NET DEBT AND NET WORTH, BILLIONS OF DOLLARS

	1946	1950	1960	1970	1980
Assets					
Tangible	170.1	112.2	207.3	301.0	727.0
Reproducible Assets	162.6	102.6	188.9	264.6	607.6
Residential Structures	2.6	2.2	3.2	5.7	20.6
Nonresidential Structures	33.1	39.1	60.8	100.2	251.6
Equipment	75.4	34.4	65.6	95.3	202.4
Inventories	51.6	26.9	59.2	63.4	132.9
Land	7.5	9.5	18.4	36.4	119.5
Financial	74.2	93.9	117.8	218.0	706.7
Currency, Demand + Time Deposits	8.8	9.7	12.8	18.0	33.9
Gold	20.7	22.8	17.8	12.0	152.3
U.S. Government Securities	24.3	21.3	28.5	63.6	123.4
Treasury Issues	24.3	21.3	28.5	63.4	114.2
Agency Issues	0.0	0.0	0.0	0.2	9.2
Mortgages	2.1	2.8	10.9	31.5	126.2
Other Loans	8.1	16.0	25.0	64.8	199.7
Taxes Receivable	8.2	16.5	12.7	5.7	9.3
Miscellaneous Assets	2.0	4.7	10.0	22.2	62.0
Total Assets	244.3	206.0	325.1	519.0	1,433.7
Liabilities					
Treasury Currency + SDR Certificates	2.4	2.4	2.7	6.0	13.6
Demand Deposits + Currency	30.4	28.2	30.6	52.0	121.5
Bank Reserves + Vault Cash	18.2	19.9	20.4	31.2	47.3
Credit Market Instruments	234.3	219.5	240.2	324.9	834.6
Savings Bonds	44.4	50.0	46.5	53.1	68.4
Other Treasury Issues	188.8	168.1	185.9	232.4	617.8
Agency Issues	1.2	1.8	7.8	39.3	148.4
Insurance, Retirement Reserves	8.0	12.7	20.5	34.9	85.5
Miscellaneous Liabilities	6.8	7.4	10.9	21.8	51.9
Total Liabilities	300.1	290.0	325.2	407.7	1,154.3
Net Debt[a]	225.9	196.2	207.4	252.7	447.5
Net Worth	−55.9	−84.0	−0.1	48.3	279.4

[a] Total liabilities minus financial assets.

ment cost, which we will take as market value, of $608 billion.[5] We estimate land owned by the federal government at $119 billion. Total tangible assets at the end of 1980 were thus $727 billion.

We add $707 billion of financial assets to reach a total assets figure at the end of 1980 of $1,434 billion. We subtract liabilities of $1,154 billion. The result, shown in Table 3, is a positive net worth of $279 billion.

The contrasting results for the series of gross Treasury debt and total liabilities on the one hand, and net debt and net worth on the other, are striking. Gross Treasury debt of $259 billion in 1946 had risen only to $321 billion by 1965, but then increased at an accelerating rate to the $930 billion by the end of 1980 (and now of course well over $1 trillion). Net debt as late as 1966 was, at $223 billion, still below the 1946 high point of $226 billion. Net debt climbed thereafter to a total of $448 billion by the end of 1980.[6]

---

[5] Of the $251 billion of federal nonresidential structures in 1980, "conservation and development" (including dams and power plants) amounted to $101 billion; military was $82 billion; industrial was $23 billion; highways and streets were $14 billion; hospitals were $8 billion; other buildings and structures were $23 billion. Of the $202 billion of federal equipment, $167 billion was military and $35 billion was nonmilitary. Residential structures were valued at $21 billion and inventories at $133 billion.

The propriety of taking the BEA replacement costs as estimates of market value is open to question, but to go beyond this would entail an effort, with still questionable results, considerably greater than we have been able to undertake at this time.

[6] Looking at what would for a private business be the bottom line, the net worth figures tell a different story, although a story that would hardly be unfamiliar to any successful large corporation. For all the while that gross Treasury debt, or even the difference between total

TABLE 4—THE REAL VALUE OF FEDERAL DEBT

	Billions of 1972 Dollars					1972 Dollars	
	Gross Public Debt	Net Debt	Budget Surplus (Deficit−)	Net Revaluation of Net Debt	Change in Net Debt	Net Debt per Capita	Change in Net Debt per Capita
1946	539.2	470.1	8.0	−83.6	–	3,384	–
1950	459.1	350.8	17.2	−31.8	44.0	2,326	−338
1960	421.4	299.1	4.4	8.7	5.8	1,679	6
1970	415.3	269.9	−15.3	1.7	15.4	1,338	60
1975	445.9	278.9	−55.2	−2.2	57.3	1,320	259
1976	481.3	318.2	−40.2	−4.2	39.3	1,482	162
1977	498.6	320.5	−33.2	−37.9	2.3	1,482	−0
1978	503.4	296.7	−19.5	−51.1	−23.8	1,364	−118
1979	496.6	244.8	−9.1	−71.4	−51.9	1,120	−244
1980	495.9	238.6	−34.5	−42.9	−6.2	1,078	−42

B. *Deficits and Real Changes in Net Debt*

While the federal government reported budget deficits totalling $336 billion from 1947 to 1980, the market value of net federal debt grew by only $222 billion from the end of 1946 to the end of 1980 (Table 3). And over that period, the real market value of that debt (in 1972 dollars) declined by $231 billion (Table 4). Even the usual measure of Treasury debt known as the gross public debt was no larger in 1980 in real terms than in 1946. In fact, it declined in 1972 dollars from $539 billion to $496 billion. This reduction in real gross debt reflects large negative net revaluations. Higher interest rates lowering the market value of debt, higher prices of gold increasing the value of that government asset and, particularly, general inflation, combined further to cut the real market value of the *net* debt virtually in half since the end of World War II.

This suggests that our conventional measure of the budget deficit is particularly misleading in periods of significant inflation, and high and fluctuating nominal interest rates. Federal government net interest payments of $53.1 billion in 1980 were only modestly less than the $61.4 billion deficit (on National Income Accounts). With interest rates on Treasury securities rising from about 11.5 percent to over 14 percent, net interest payments in 1981 came to $71.9 billion, well exceeding a $60.0 billion reported deficit.[7] (For 1982, however, while net interest payments rose to $85.0 billion, the deficit on national income account soared to $149.5 billion.)[8] But measures of real interest, taking into account changes in the real market value of the federal debt, would have been small fractions of nominal payments, or actually negative. Substitution of real for nominal interest costs in budget expenditures would indicate that the budget was close to balanced in 1980, a result consistent with the changes in the real value of the debt (Table 5).[9]

---

liabilities and total financial assets which we have been calling net debt, was growing, so was net worth. By our measures, net worth was a negative $56 billion at the end of 1946, as shown in Table 3. It became even more negative, finally reaching a lower bound of minus $92 billion at the end of 1949. But, by the end of 1956, the federal government was at least briefly in the black, with a net worth of $8 billion. Net worth was then trivially negative, hovering around zero, for the years 1960–62, before it again began to rise. It dipped in the 1974–75 recession and its wake, but came up very sharply in the next five years, when gross Treasury debt was soaring dramatically. We measure federal net worth at the end of 1980 at $279 billion. Our federal government, like many individuals and business that borrowed heavily in an inflationary period, came out pretty well with its investments in tangible assets—and in gold.

[7,8] *Economic Indicators*, May 1983, p. 34.

[9] Of course, the population has been growing over the last 35 years. Our ultimate measure of the trend in debt then may well be the real net debt per capita. This figure declined (in 1972 dollars) from $3,384 at the end of 1946 to $1,078 at the end of 1980, as also shown in Table 4.

(over)

TABLE 5—RECONCILIATION OF FEDERAL BUDGET SURPLUS (DEFICIT) AND CHANGE IN REAL NET DEBT

(1)	Surplus (Deficit −) on National Income Accounts (2)	Interest Rate Effects (Par-to-Market Adjustment) (3)	Price Effects (Nominal to real adjustment) (4)	Adjusted Surplus (Deficit −) (5)	Off-Budget Items[a] (6)	Net Revaluation on Gold (7)	Decrease (Increase −) in Real Net Debt in Current Dollars (8)
1976	−53.1	−12.5	21.2	−44.5	−4.2	−3.3	−52.0
1977	−46.4	16.9	29.3	−.2	−9.0	6.0	−3.3
1978	−29.2	18.6	43.3	32.7	−8.7	11.8	35.7
1979	−14.8	1.7	44.1	31.0	−9.8	63.6	84.9
1980	−61.2	12.6	55.9	7.3	−2.1	6.0	11.2

Note: Column 5 is total of cols. (2), (3), and (4); Column (8) is total of cols. (5), (6), and (7).

[a] This is the sum of insurance credits to households, federal mineral sales, credit agency surplus, net revaluations on foreign exchange and miscellaneous assets, and the Flow of Funds statistical discrepancy.

A proper measure of the government's current surplus is the reduction in real value of government net debt or, consistent with the Haig-Hicks definition of income, the amount of additional expenditures that can be made while keeping net debt intact. If there were no net revaluations of existing financial assets and liabilities, this definition would correspond to conventional measures of the deficit. Net revaluations have in fact generally been considerable, however, as we have just noted, and all the more so in periods of significant inflation and high and fluctuating nominal interest rates.

As with the debt, therefore, we make two adjustments to the deficit, a par-to-market adjustment and a nominal-to-real adjustment. The first of these is the change in net debt due to changing market rates of interest, which we term the interest effect. With generally rising interest rates, this effect has been on balance positive in the postwar era, thus reducing the adjusted deficit.

The second and usually larger adjustment, incorporating what we denote as the price effect, is the change in the real value of the net debt due to changes in the price level. It has of course become of major importance with the higher inflation of the 1970's.

By reducing our measure of the deficit by the amount of decrease in the real value of existing net government debt, we are in effect including the so-called "inflation tax" in budget revenues. Since inflation (and expected inflation) forces borrowers to offer higher nominal interest rates to compensate for the inflation tax on holders of securities, the price-effect adjustment may alternatively be viewed as a conversion of interest expenditures from nominal to real.

In Table 5 we add the interest and price effects to present an adjusted measure of the federal budget surplus or deficit. The conventional and adjusted budgets prove quite different. The officially reported deficit of $61.2 billion in 1980, for example, is converted to a $7.4 billion surplus when price and interest effects are considered. And the adjustments have not been constant in amount; in earlier years, when inflation was lower, they were decidedly smaller.[10]

II. The High-Employment Budget

A. *The Official High-Employment Budget Surplus and Relevant Adjusted Measures*

The high-employment budget, currently calculated for an unemployment level of 5.1

Real debt per capita was thus at the end of 1980 less than one-third of what it was at the end of 1946.

[10] The adjusted surplus differs from the real change of government net debt, expressed in current dollars, by the amounts of the off-budget items and net revaluations of gold. We have excluded net revaluations on gold from the adjusted surplus on the principle that an increase in the value of government gold holdings, unlike an increase in the value of government holdings of private debt, does not decrease private net worth. It may indeed be argued, rather, that an increase in the value of gold will be perceived by the private sector as increased backing for government debt and thus as increased private wealth. Similar considerations apply to the off-budget items.

percent, is generally understood to provide a first-order measure of the thrust of federal fiscal policy on aggregate demand. Increases in the high-employment budget surplus or decreases in the deficit are viewed as contractionary. The tendency of the high-employment budget to be in deficit in recent years, in contrast to those before 1966, has perhaps contributed to the view that, despite generally increasing inflation, the federal budget has been, perversely, expansionary. The depth and severity of the recession of 1981–82 have then been attributed to the great potency of a tight and tough monetary policy which has overbalanced an easy fiscal policy.

These views, at least of fiscal policy, are fundamentally flawed. In particular, in a period of significant and rising inflation, the traditional high-employment budget, as usually measured, is no longer an unbiased measure of fiscal thrust. It in fact measures nominal rather than real impacts on income, and confuses nominal flows with changes in real stocks.

For the high-employment budget to measure correctly the thrust of federal fiscal policy, it must measure correctly its contribution to aggregate demand. We abstract here from not unimportant questions relating to the balanced budget multiplier, and to the mix among both expenditures and receipts. We abstract as well from the question of monetization of the debt or possible interaction of fiscal and monetary policy. These issues aside, a relevant high-employment deficit must add to correctly measured private income and wealth.

It is clear, for example, that a "deficit" which involves expenditures by the Treasury to purchase existing securities held by the public would not be a relevant deficit. Neither, indeed, would a deficit used to finance purchase of existing assets which merely involves a change in private portfolios—the substitution of government bonds for the assets sold to the Treasury. Similarly, a surplus brought on by the sale of public lands to the private sector does not change private wealth holdings—except possibly to raise them if the sales are at prices below market value. And a deficit that does not increase the net debt of the government does not increase the net income or net worth of the private sector, and hence does not per se have an expansionary effect on aggregate demand.[11]

Perception of the gap in the relation between measured deficits and changes in net debt of the government, or net worth of the private sector, points to the major effect of inflation on a relevant high-employment budget. For inflation and changing rates of inflation quite break whatever links have existed between budget deficits and real income or changes in net worth of the private sector.

A measure of the real actual surplus or deficit may be viewed as essentially the sum of three components: 1) the nominal surplus or deficit as currently measured; 2) an adjustment for changes in market value of government financial assets and liabilities due to changing market rates of interest; and 3) changes in the real value of net debt due to changing general price levels incident to inflation. An identical or analogous set of adjustments is appropriate for the high-employment budget surplus or deficit.

It should be understood that the inflation adjustments deal with rates of change in the price level and rates of change in the rate of change in the price level. Price level effects are included in the definition and measure of the high-employment budget currently. Frank de Leeuw and Thomas Holloway (1982) present in some detail the sophisticated measures utilized in tabulating automatic effects of inflation and higher price levels on the high-employment budget surplus or deficit as defined by the Bureau of Economic Analysis. Higher price levels and inflation cause more than proportionate increases in corporate taxes, and in personal taxes as long as the

---

[11] We do view increases in net debt as, *ceteris paribus*, increasing private perceptions of financial wealth and hence private demand for goods. With Ricardo himself we hence reject the Ricardo "equivalence theorem," by which the private sector would perceive increases in the present value of future tax liabilities equal to any current increases in government debt. And whether increased real government budget deficits would result in "crowding out," with no net increase in total output, as argued by William Fellner (1982), we would suggest would depend upon how close the economy is in fact to full employment. With ample unutilized resources, it is hard to see why fiscal stimulus must prove abortive.

TABLE 6—HIGH-EMPLOYMENT BUDGET SURPLUS (DEFICIT) AS PERCENT OF *GNP*, OFFICIAL AND ADJUSTED

	High-Employment Surplus (Deficit −)		
	Official	Adjusted for Price Effects	Adjusted for Price and Interest Effects
1955	1.30	2.81	3.71
1956	1.87	3.83	4.63
1957	1.37	2.47	1.26
1958	0	.93	2.21
1959	1.11	2.09	2.90
1960	2.39	2.84	.98
1961	1.35	1.99	2.44
1962	.53	1.29	.91
1963	1.24	1.79	2.26
1964	.17	.78	.73
1965	.13	.98	1.42
1966	−.74	.34	.09
1967	−1.89	−.89	−.38
1968	−1.26	.06	.20
1969	.52	1.95	2.68
1970	−.46	.77	−.64
1971	−1.05	.12	−.21
1972	−1.02	.02	.42
1973	−.72	.89	1.13
1974	−.02	2.16	1.97
1975	−1.88	−.37	−.55
1976	−1.01	.22	−.52
1977	−1.13	.39	1.27
1978	−.70	1.30	2.16
1979	−.09	1.74	1.81
1980	−.81	1.31	1.63
1981	.15	1.92	1.97

rate structure is progressive and not indexed. While many expenditure programs (Social Security most conspicuously) are also inflated by higher prices, and interest payments, as observed, are increased by the higher nominal rates associated with higher inflation, the effect of inflation on balance is to increase the high-employment budget surplus or decrease the deficit.

But these effects of inflation on the high-employment budget surplus as defined are quite another matter than the effect of inflation on the real high-employment budget. For the greater the rate of inflation, the greater is the decline in the real value of net private holdings of government debt or, put another way, with real interest costs substituted for nominal interest payments, the greater is the real high-employment budget surplus (or the less the deficit).[12]

We have hence undertaken to calculate a high-employment budget that, by correcting for these inflation effects, comes closer to measuring the real surplus or deficit and consequent thrust of fiscal policy on aggre-

[12] The high-employment budget, while aimed at eliminating the cyclical effects on the actual budget, is of course not exogenous to what is going on in the economic system. To consider it a measure of fiscal thrust is not to view it as entirely at the discretion of the fiscal authorities. And just as an easy monetary policy lowering Treasury interest payments, or inflation as a consequence of external supply shocks, will move the official high-employment budget to surplus, so will *tight* money and inflation move our price- and interest-adjusted high-employment budget further to surplus, with no independent government action.

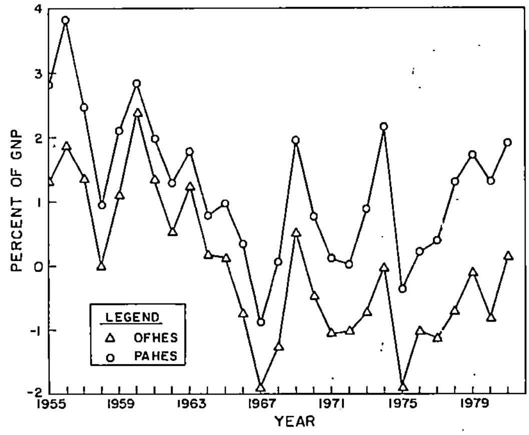


FIGURE 1. OFFICIAL AND PRICE-ADJUSTED HIGH-EMPLOYMENT SURPLUS

gate demand. Applying our calculations of net revaluations on actual net federal debt, we have adjusted the high-employment budget surplus series presented by de Leeuw and Holloway for the years 1955 through 1981. The results shown in Table 6, and partially in Figure 1, are dramatic. Inflation and rates of interest were low and relatively steady in the early years of the period prior to escalation of our military involvement in Vietnam. Corrections to the high-employment budget surplus are hence small in those early years. In the years from our heavy Vietnam involvement on, however, when the high-employment budget as well as the actual budget moved substantially to deficit, the effect of our corrections is dramatic. The entire perceived trend in the direction of fiscal ease or expansion is eliminated or reversed. The adjusted high-employment budget surplus, as a percent of *GNP*, for every year from 1977 through 1981 was, except for the tax-surcharge year of 1969 and the oil-price-shock year of 1974, higher than the surplus of any single year going back to 1965. With similar exceptions, the surplus adjusted only for price (and not interest) effects was higher in every year from 1978 to 1981 than that of any single year back to 1963.

This suggests a need for some significant rewriting of recent economic history. Inflation can hardly be ascribed to excess demand associated with increasing fiscal ease and stimulus if, at least by the high-employment budget measure, appropriately defined, there was no such movement to fiscal ease.[13] Some explanation of sluggishness in the economy in the last several years, climaxed by the current recession, may then be found in a relatively tight fiscal policy, as measured by the adjusted high-employment budget surplus, as well as in the widely perceived role of monetary policy.

B. *Empirical Relations*

Movements in real economic activity in the United States have in fact been negatively related to the high-employment budget surplus. This is confirmed dramatically in relations between percent changes in real *GNP* and unemployment on the one hand, and lagged values of the high-employment budget surplus as a percentage of *GNP* on the other.

Recognizing 1967 as something of a watershed inaugurating accelerating inflation and subsequently slower rates of growth and higher unemployment, we have divided our observations into two groups: 1956–66 and 1967–81 (or 1982). Looking first at the official high-employment budget surplus (*OFHES*), we note in equations (1) and (4) of Table 7 that it was clearly negatively

---

[13] That federal government expenditures as a share of *GNP* rose, from 17.0 percent in 1955 to 23.4 percent in 1981 (while receipts rose from 18.2 to 21.4 percent), is also not evidence of an expansionary fiscal policy. Some part of the expenditure increase involves the increase in nominal interest payments which would be viewed as akin to a purchase of government liabilities of real debt. Federal expenditures net of net interest paid increased only from 15.9 to 21.0 percent of *GNP*. More generally, the balanced budget multiplier will apply to government expenditures for goods and services, not to the transfer payments financed by taxes. The share of goods and services of *GNP* purchased by the federal government declined sharply from 11.1 percent in 1955 to 7.8 percent in 1981. Only when state and local expenditures are included is the picture altered. Total government expenditures for goods and services rose from 18.8 percent of *GNP* in 1955 to 20.3 percent in 1981. In the latter year, federal grants in aid to state and local governments came to 3.0 percent of *GNP*. (Percents calculated from dollar figures in *Economic Report of the President*, February 1982, pp. 233 and 320; *Economic Indicators*, May 1983, pp. 1 and 34; and, for 1955 net interest, *National Income and Product Accounts, 1929–74*, p. 96.)

TABLE 7—HIGH-EMPLOYMENT BUDGETS AND CHANGES IN REAL *GNP* AND UNEMPLOYMENT

Equations	Measure of *HES*	Dependent Variable (Y)	Regression Coefficients[a]				$\hat{R}^2$	D-W
			1956–66		1967–81			
			b_{01}	b_{11}	b_{02}	b_{12}		
(1)	*OF*	%Δ*GNP*	5.678	−1.837	.519	−2.896	.451	1.61
			(.930)	(.733)	(.750)	(.730)		
(2)	*PA*	%Δ*GNP*	7.189	−1.730	4.432	−2.318	.595	2.02
			(1.089)	(.499)	(.496)	(.439)		
(3)	*PIA*	%Δ*GNP*	4.838	−.505	−1.055	−1.561	.291	2.12
			(1.240)	(.508)	(.625)	(.470)		
(4)	*OF*	%Δ*UN*	−.881	.792	1.281	1.258	.399	2.25
			(.437)	(3.44)	(.352)	(.343)		
(5)	*PA*	%Δ*UN*	−1.182	.569	−.384	.955	.403	2.31
			(5.93)	(.272)	(.270)	(.250)		
(6)	*PIA*	%Δ*UN*	−.116	−.182	.333	.493	.085	2.01
			(.319)	(.259)	(.632)	(.240)		
			1956–66		1967–82			
(2′)	*PA*	%Δ*GNP*	7.189	−1.730	4.450	−2.524	.638	2.01
			(1.068)	(.490)	(.504)	(.438)		
(5′)	*PA*	%Δ*UN*	−1.182	.569	−.389	1.018	.462	2.31
			(.594)	(2.272)	(.268)	(.233)		

Notes: *HES* = high-employment budget surplus; *OF* = official; *PA* = price adjusted; *PIA* = price and interest adjusted; %Δ*GNP* = percent change in real Gross National Product; %Δ*UN* = percentage point change in unemployment.

$$Y_t = b_{01} + b_{11} HES_{t-1} \qquad t = 1955\text{–}66$$

$$Y_t = b_{02} + b_{12} HES_{t-1} \qquad t = 1967\text{–}81, 1982$$

[a] Standard errors are shown in parentheses.

related to subsequent growth in real *GNP* and positively related to subsequent changes in employment in both periods, apparently more sharply so (although the differences are not statistically significant) in the later period. Also note that in the earlier period, a balanced official high-employment budget would "predict" a 5.7 percent growth in real *GNP* and an annual decline of 0.9 percentage points in unemployment. In the later period, however, a balanced official high-employment budget would predict only a 0.5 percent growth in real *GNP*, along with a 1.3 percentage point annual increase in unemployment. Each percentage point increase in the official high-employment budget surplus (measured as a percent of *GNP*) is associated with a reduction of 2.9 percentage points in the rate of growth of real *GNP* and an increase of almost 1.3 percentage points in unemployment.

Why is a balanced official high-employment budget now associated with a stagnant economy and increasing unemployment? A clue to the answer is found in the relations involving the price- or inflation-adjusted high-employment budget, shown in equations (2) and (5) of Table 7 and remarkably sharply in Figures 2 and 3. We note first, from the values of b_{02} in equations (2) and (5), that a balanced, *price-adjusted* high-employment budget in the later periods predicts a 4.4 percent growth in real *GNP* and a 0.4 percent decline in unemployment. But, recall from Table 6 that in 1981, when the official high-employment surplus was 0.15 percent of *GNP*, the high-employment surplus adjusted for price effects was 1.92 percent of *GNP*—a difference of 1.77 percentage points. Our relations suggest that if the official high-employment budget had been in deficit by that much, real growth would

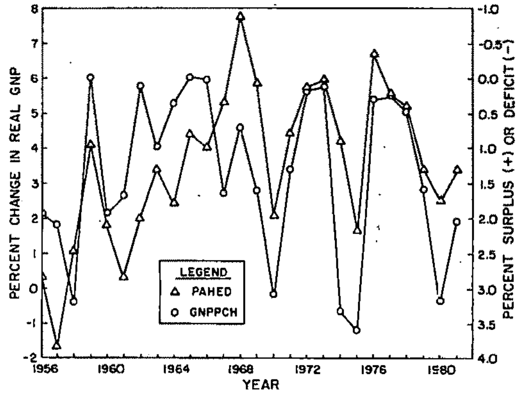


FIGURE 2. LAGGED PRICE-ADJUSTED H-E DEFICIT AND CHANGE IN *GNP*

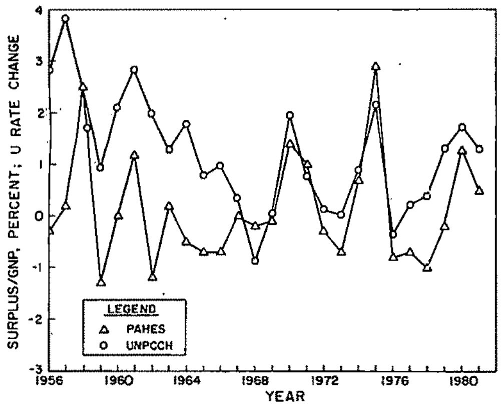


FIGURE 3. LAGGED PRICE-ADJUSTED H-E SURPLUS AND CHANGE IN UNEMPLOYMENT

have risen to 5.6 percent and unemployment would have declined by 0.9 percentage points.[14]

What is more, the coefficients of determination in the relations involving the price-adjusted surplus, particularly the value of .595 in the *GNP* change equation, suggest that the inflation effect does at least as well, and possibly better, in capturing the impact of fiscal thrust than the components of the official high-employment budget surplus. This would indeed be consistent with the hypothesis that in the official unemployment budget relation we have an errors-in-variable problem. The true relation then involves the price-adjusted high-employment budget. In the earlier period, when inflation was small, the relations were fairly similar. In the later period with high inflation, the differences in the relations cannot safely be ignored.

Our relations involving one-year lagged, adjusted high-employment budget surpluses became poorer, however, when the corrections relating to par-to-market adjustments of the value of the debt were incorporated. As shown in equations (3) and (6), estimated parameters retained the hypothesized signs in all cases, except the earlier period for the unemployment equation, but absolute values of the estimates, levels of significance, and coefficients of determination were less. These results may stem in part from the high variability of interest rates and consequent par-to-market corrections which may cause private creditors to view interest effects to a considerable extent as transitory. A par-to-

---

[14]
$$b_{11}+b_{12}(-1.77)=.519-2.896(-1.77)$$
$$=5.644 \text{ in } (1),$$
$$b_{11}+b_{12}(-1.77)=1.281+1.258(-1.77)$$
$$=-0.946 \text{ in } (4).$$

These are, of course, disequilibrium relations. We do not suggest on the basis of (2) that a balanced price-adjusted high-employment budget would generate a permanent growth in real *GNP* of 4.432 percent or, on the basis of (5), permanent rates of decline in unemployment of 0.384 percentage points per year. Indeed, one might infer that it would prove impossible to maintain a balanced, adjusted high-employment budget indefinitely. For as unemployment reached truly full-employment levels, further stimuli to demand would increase the rate of inflation until the price-adjusted high-employment budget were sufficiently in surplus to be consistent with no further changes in the rate of unemployment. By way of illustration, without taking the specific numbers too seriously, equation (5) suggests that the rate of unemployment would be constant with a price-adjusted high-employment surplus equal to 0.402 percent of *GNP*. This would imply a 3.5 percent per annum growth in real *GNP*.

Similarly, equation (4) indicates that zero change in unemployment would require an official high-employment deficit equal to 1.018 percent of *GNP*. This would also imply real per annum growth in *GNP* of some 3.5 percent.

P3534

TABLE 8—HIGH-EMPLOYMENT BUDGETS AND CHANGES IN REAL *GNP* AND UNEMPLOYMENT

Equations	Measure of *HES*	Dependent Variable (Y)	Regression Coefficients[a]							
			1957–66			1967–81				
			b_{01}	b_{11}	b_{21}	b_{02}	b_{12}	b_{22}	$\hat{R}^2$	$D\text{-}W$
(7)	*OF*	%Δ*GNP*	5.634	−1.798	.104	.522	−2.896	.003	.399	1.63
			(1.296)	(.799)	(.865)	(.976)	(.777)	(.726)		
(8)	*PA*	%Δ*GNP*	7.794	−1.504	−.488	4.301	−2.349	.236	.565	2.04
			(1.403)	(.608)	(.621)	(.590)	(4.87)	(4.95)		
(9)	*PIA*	%Δ*GNP*	7.903	−.281	−1.554	4.474	−1.494	−.648	.603	1.97
			(1.265)	(4.27)	(.395)	(.531)	(.359)	(.366)		
(10)	*OF*	%Δ*UN*	−.897	.808	.040	1.497	1.265	.278	.361	2.33
			(.600)	(.370)	(.401)	(.452)	(.360)	(.336)		
(11)	*PA*	%Δ*UN*	−1.710	.498	.353	−.455	.938	.126	.403	2.40
			(.735)	(.320)	(.327)	(.311)	(2.36)	(.260)		
(12)	*PIA*	%Δ*UN*	−1.289	−.244	.790	−.340	.457	.347	.484	2.15
			(.648)	(.219)	(.202)	(.272)	(.184)	(.187)		
			1957–66			1967–82				
(9′)	*PIA*	%Δ*GNP*	7.903	−.281	−1.554	4.513	−1.623	−.764	.633	2.05
			(1.240)	(.418)	(.387)	(.548)	(.348)	(.361)		
(12′)	*PIA*	%Δ*UN*	−1.289	−.244	.790	−.355	.504	.347	.521	2.20
			(.656)	(.221)	(.205)	(.273)	(1.95)	(.181)		

Notes: See Table 7.

$$Y_t = b_{01} + b_{11}HES_{t-1} + b_{21}HES_{t-2} \qquad t = 1957 \text{ to } 1966.$$

$$Y_t = b_{02} + b_{12}HES_{t-1} + b_{22}HES_{t-2} \qquad t = 1967 \text{ to } 1981, 1982$$

[a] See Table 7, fn. a.

market correction based on market values of securities on December 31 may not be promptly reflected in aggregate demand of the subsequent year.

We have therefore estimated relations in which changes in *GNP* and employment are functions of both one- and two-year lags of the high-employment budget surplus. The results, presented in Table 8, are again dramatic.

The effects of the official and price-adjusted high-employment surplus are essentially felt with only a one-year lag, as indicated in equations (7), (8), (10), and (11). In the *PIA* (price- and interest-adjusted) equations, adjusted coefficients of determination ($\hat{R}^2$) are uniformly lower with the extra lagged variables in the *OFHES* and *PAHES* equations, while they are markedly higher in *PIAHES* equations. Indeed, the two-year distributed lag equations for the *PIAHES* (the fully adjusted high-employment budget surplus) show, at least trivially, the best fit of all, with adjusted coefficients of determination of .603 and .484 for the percent change in *GNP* and the percentage point change in unemployment, respectively.[15]

---

[15] The distributed-lag, price- and interest-adjusted equations (9) and (12) suggest that a balanced, adjusted budget would, after two years, generate per annum rates of growth of real *GNP* of 4.5 percent and annual declines of unemployment of 0.34 percentage points. A steady state in unemployment would require an adjusted budget in surplus by 0.423 percent of *GNP*. This would in turn imply a rate of growth of real *GNP* of 3.57 percent.

As this paper was completed, preliminary estimates of changes in real *GNP* and unemployment for 1982 became available. We used them in estimating equations (2′) and (5′), respectively, in Table 7, and (9′) and (12′) in Table 8. They improved the fit and confirmed the relations estimated for the years through 1981. The originally estimated equations indeed went a considerable way in predicting the deepening recession in 1982. Real *GNP* in 1982 was down 1.7 percent after a 1.9 percent rise in 1981. Equation (1), based on the official high-employment budget surplus, would have predicted *GNP* growth in 1982 of 0.08 percent. Equation (2), based on the price-adjusted surplus would have come a bit closer in catching the slowdown, forecasting a de-

III. Summary and Conclusions

We have noted first that official measures of the federal debt and budget deficits are misleading by any of several reasonable standards. Gross public debt figures ignore financial asset accumulation as well as the real assets which have contributed to a growing government net worth. Budget flows have failed to distinguish between current and capital accounts, and measures of surplus and deficit have been inconsistent with changes in the real value of net debt.

Adjustment for the changing relations between the par and market value of government financial assets and liabilities and further adjustments for the effects of inflation on the real value of net debt reveal a very great secular *decline* since the end of World War II in the explicit net claims (excluding Social Security and other future retirement benefits) of the private sector on the federal government. This suggests that by relevant economic measures, consistent with changes in the real value of the debt, the federal budget may properly be viewed as more frequently in surplus than in deficit.

Similar adjustments applied to the official high-employment budget surplus or deficit alter sharply the conventional picture of recent fiscal policy, particularly in periods of substantial inflation. An appropriately adjusted high-employment budget turns out to have been not in deficit in recent years, as usually supposed, but in considerable surplus. The view that fiscal policy has generally been too easy and overstimulatory is contradicted. Even with the new tax cuts, as late as 1981 the adjusted high-employment budget was very substantially in surplus and a relatively contractionary force in the economy.

Empirical confirmation of this view may be found in simple, lagged relations between changes in real *GNP* and changes in unemployment on the one hand, and the official and adjusted measures of the high-employment budget surplus or deficit on the other. The official high-employment budget surplus and, even more clearly, the inflation-adjusted and fully adjusted high-employment budget surpluses were substantially and significantly negatively related to real growth and positively related to changes in unemployment.[16] A return to recovery rates of growth of *GNP* and reduction in unemployment, at the rates of inflation and interest rates through 1981, would call for balanced, inflation-adjusted high-employment budgets, or much larger deficits in the official high-employment budget.[17]

This suggests two important correctives to widespread views of fiscal and monetary policy. First, the 1981–82 recession cannot be properly interpreted as the triumph of all-powerful monetary constraints over relatively ineffective fiscal ease. Without denying the significant effects of a tight money policy in depressing the economy, it is necessary to point out that the verdict is not in on what might have happened if tight money had been pitted against real fiscal stimulus.

Second, the absence of real fiscal stimulus —and indeed the presence of some measure of fiscal tightness—makes clear the extent of the overkill of monetary restraint which contributed to the highest unemployment levels in almost two generations. Those who acqui-

---

cline of 0.02 percent. Equation (9) forecast *GNP* growth of 0.47 percent.

The one-year lag equations predicted about half of the increase in unemployment which occurred from 1981 to 1982. Might we attribute the remaining half to monetary policy and the gathering force of cyclical decline?

[16] Our estimates are of course essentially of reduced-form relations. We should hope that the results are sufficiently suggestive to induce application of our adjustments in full macroeconometric models. This might permit clarification of the extent to which price and interest movements are affecting consumption and investment directly via their effects on net debt or via associated real interest rates which may in turn be related to monetary policy. It might also illuminate the extent to which federal surpluses and deficits have been offset at the state and local level.

[17] Presently anticipated official high-employment deficits in the future are another matter. As indicated in Appendix B and Table A2, current Congressional Budget Office projections of actual deficits and rates of interest, inflation, and unemployment suggest adjusted high-employment budget deficits of 1.60, 1.99, and 2.55 percent of *GNP* in 1986, 1987, and 1988, respectively, implying an extremely expansionary but unsustainable fiscal policy.

esced in tight money as "the only game in town" to slow a presumably overheated inflationary economy were wrong on two counts. The absence of increasing fiscal thrust reinforces the inference that the inflation came on the side of supply shocks rather than excess demand. And the strong-willed rejection of accommodative monetary policy, rather than balancing budget excesses, offered a near-lethal combination of monetary and fiscal contraction.

APPENDIX

A. *Total Government: Federal, State, and Local*

Total liabilities of state and local governments grew from $16 billion at the end of 1946 to $296 billion at the end of 1980, as shown in Table A1. For that same period, however, their financial assets grew from $12 billion to $205 billion. Net debt was close to zero in 1946, and had increased only moderately to a peak of $120 billion at the end of 1976 before turning down.

When we take into account tangible assets —highways, water and sewer systems, public buildings, land, and the like—we find that net worth of state and local governments, always positive, grew enormously. From 1946, when it stood at $82 billion, it rose to $1,605 billion by the end of 1980.

We may look at the sum of the balance sheets for federal, state, and local governments to find the total government picture. There we note that while liabilities grew by over a trillion dollars, from $316 billion in 1946 to $1,450 billion by the end of 1980, financial assets grew over that same period from $87 billion to $912 billion. Thus net debt grew only from $230 billion to $538 billion.

Total government tangible assets, however, grew by over $2 trillion, from $255 billion at the end of 1946 to $2,423 billion at the end of 1980. Thus, the value of total assets of government grew by almost $3 trillion, from $342 billion at the end of 1946 to $3,335 billion at the end of 1980. The net worth of total government also grew by nearly $2 trillion, from $26 billion at the end of 1946 to $1,884 billion at the end of 1980.

B. *Projected Future High-Employment Budgets and their Effects*

With massive tax cuts and increases in military spending dwarfing reductions in nonmilitary government outlays, fiscal policy took a sharp turn towards stimulus in 1982. This was measured in the official high-employment budget, which moved from a surplus equal to 0.15 percent of *GNP* in 1981 to a deficit of 1.08 percent of *GNP* in 1982. Congressional Budget Office calculations consistent with projected budget deficits from 1983 to 1988 indicate that the official high-employment deficit will grow further, to 1.75 percent of *GNP* in 1983, reaching 4.14 percent in 1988.

Utilizing Congressional Budget Office projections of movements in the *GNP* implicit price deflator and Treasury bill rates, along with Wharton Annual Model projections of interest rates on U.S. government notes and bonds, and a future series for federal debt developed from forecasts of actual deficits, we have also been able to calculate the high-employment budget adjustments for price and interest effects. To do so, we constructed market-to-par indices for future years by extrapolating relations between these indices and the relevant interest rates over the earlier period, relations which fit quite well, and then applied these and the projected price changes to projections of financial assets and liabilities assumed to have the same distribution as in 1981.

The results shown in Table A2 for the years 1982–88 indicate a pattern quite different from the corresponding data in Table 6 for earlier years. Where our adjustments for price and interest effects in the recent period of high inflation and rising interest rates showed adjusted budgets moving more sharply to restraint, the impact of the adjustments now is in the direction of relatively greater stimulus. While the official high-employment budget moved from 0.15 percent in surplus in 1981 to 1.08 percent in deficit in 1982, as noted above, the adjusted high-employment budget, with declining rates of inflation and falling interest rates, moved much more. The high-employment budget adjusted only for price effects went from 1.92 percent in surplus to a virtual balance, of

TABLE A1—ASSETS, LIABILITIES, AND NET WORTH, BILLIONS OF DOLLARS, FEDERAL, STATE AND LOCAL, AND TOTAL GOVERNMENT

	Assets					
Year (1)	Tangible (2)	Financial (3)	Total (4)	Total Liabilities (5)	Net Debt (6)	Net Worth (7)
Federal						
1946	170.1	74.2	244.3	300.1	225.9	−55.9
1950	112.2	93.9	206.0	290.0	196.2	−84.0
1960	207.3	117.7	325.0	325.2	207.5	−.2
1970	301.0	218.0	519.0	470.7	252.7	48.3
1975	433.6	359.3	793.0	720.0	360.7	72.9
1976	462.9	397.0	859.9	829.0	432.0	31.0
1977	509.7	428.5	938.2	890.6	462.1	47.6
1978	571.8	504.9	1,076.7	970.1	465.2	106.6
1979	647.4	632.6	1,280.1	1,049.3	416.6	230.8
1980	727.0	706.7	1,433.7	1,154.3	447.5	279.4
State and Local						
1946	85.2	12.5	97.7	16.2	3.7	81.5
1950	114.4	17.6	132.0	26.6	9.0	105.4
1960	237.2	32.8	270.0	74.2	41.5	195.8
1970	548.5	70.6	619.1	149.8	79.2	469.3
1975	977.8	121.6	1,099.4	223.4	102.3	875.5
1976	1,049.9	134.1	1,184.0	253.9	119.8	930.1
1977	1,159.7	155.7	1,315.8	273.5	117.7	1,042.3
1978	1,317.8	173.0	1,490.8	283.9	110.9	1,206.9
1979	1,504.0	189.7	1,693.7	297.7	108.0	1,396.0
1980	1,695.9	205.1	1,901.0	296.0	90.9	1,605.0
Total Government						
1946	255.3	86.7	342.0	316.3	229.6	25.7
1950	226.6	111.5	338.1	316.6	205.2	21.5
1960	444.4	150.6	595.0	399.5	248.9	195.5
1970	849.5	288.6	1,138.1	620.5	331.9	517.6
1975	1,411.4	480.9	1,892.3	943.9	463.0	948.4
1976	1,512.8	531.1	2,043.9	1,082.9	551.8	961.0
1977	1,669.4	584.2	2,253.6	1,164.1	579.8	1,089.5
1978	1,889.6	678.0	2,567.6	1,254.1	576.1	1,313.5
1979	2,151.4	822.4	2,973.8	1,347.0	524.6	1,626.8
1980	2,422.9	911.9	3,334.8	1,450.3	538.4	1,884.5

Note: Column (6) is col. (5)−col. (3); Column (7) is col. (4)−col. (5).

TABLE A2—PROJECTED HIGH-EMPLOYMENT BUDGET SURPLUS OR DEFICIT AS PERCENT OF *GNP*, OFFICIAL AND ADJUSTED

	High-Employment Budget Surplus (Deficit −)		
Year	Official[a]	Adjusted for Price Effects	Adjusted for Price and Interest Effects
1981	0.15	1.92	1.97
1982	−1.08	.01	−1.77
1983	−1.75	−.48	−1.46
1984	−2.49	−1.08	−.47
1985	−2.87	−1.37	−1.13
1986	−3.25	−1.80	−1.60
1987	−3.74	−2.26	−1.99
1988	−4.14	−2.59	−2.55

[a] Years from 1982 to 1988 calculated from Congressional Budget Office figures furnished by Frank Russek; the projected deficit for 1988 was the average of the first three quarters with the third quarter weighted twice.

0.01 percent in surplus. And the fully adjusted high-employment budget, including price and interest effects, moved almost 4 percentage points, from 1.97 percent in surplus to 1.77 percent in deficit!

It may be instructive to note what our equations with the various high-employment budgets predict for 1983, and beyond. Equations (2′), (5′), (9′), and (12′), shown in Tables 7 and 8, and new equations for the official high-employment budget estimated on data through 1982, all indicate an upturn in 1983. The official high-employment budget equation predicts the smallest increase in real *GNP*, 3.7 percent, and unemployment essentially unchanged (up about 0.1 percentage points). The price-adjusted high-employment budget, recognizing the lesser effects of inflation on the real value of the debt, indicates a growth of 4.4 percent in real *GNP* and decline of 0.4 percentage points in unemployment. Finally, the price- and interest-adjusted high-employment budget equations point to a growth in real *GNP* of some 5.9 percent and a decline in unemployment of 0.45 percentage points.

As this is written (June 1983), the Bureau of Economic Analysis has released its second-quarter "flash" estimate of real *GNP* growing at a 6.6 percent annual rate, which Martin Feldstein, Chairman of the Council of Economic Advisers, and other administration officials indicate may presage a (presumably) fourth-quarter over fourth-quarter increase of 5.5 percent.[18] The adjusted budget equations point more closely to that mark than does the equation with the official high-employment budget, but as year-over-year projections, the estimates of the adjusted equations appear somewhat high. By the time this paper is in print, the reader should have the opportunity to check the various forecasts against the actual data for 1983.

Looking further into the future, with forecasts of lower inflation and declining or steady interest rates, our adjustments do relatively little to temper projections by the Congressional Budget Office of a high-employment budget very substantially in deficit. We cannot consistently apply our projections of adjusted deficits to forecast the path of the economy because this path may diverge from the Congressional Budget Office projections on the basis of which both the official high-employment budget and our adjustments are calculated. Without taking into account the feedback which would alter the underlying Congressional Budget Office projections, our equation (12′), for example, would point to rates of growth of real output ranging from the 5.88 percent in 1983 to 8.37 percent in 1988, and an unemployment rate which would be down to 2.2 percent by 1988; applications of the official high-employment budget would indicate a 1988 unemployment rate of minus 0.8 percent!

All this has obviously ignored supply constraints. The aggregate demand which would call for these paths of output and employment must in fact be expected to generate considerably higher rates of inflation than those in the Congressional Budget Office forecasts. These in turn would affect interest rates and our price and interest adjustments to the high-employment budget. Consequently, the fully adjusted high-employment budget may in fact turn out to be less in deficit than assumed in our productions. Indeed, as increases in inflation increasingly erode the value of debt created by increases in deficits, the projected adjusted deficits may well prove unattainable.

C. *Sources and Methods*[19]

The major basic sources of data are the Bureau of Economic Analysis (BEA) of the Department of Commerce and the Flow of Funds Accounts of the Federal Reserve Board.

Data on reproducible tangible assets of government (including government enterprises) are the revised series, at current (replacement) cost, on the BEA capital stock data tape.[20] Land estimates for 1952 through 1968 originated with Grace Milgram in Raymond W. Goldsmith (1973). These were extrapolated backward and forward on the basis of total private land estimates taken from the Flow of Funds.[21]

---

[18] As reported in *The Wall Street Journal*, June 22, 1983, p. 3.

[19] A more detailed and precise description of sources and methods, including construction of the various indices developed, is available from the authors on request.

The Flow of Funds data on financial assets and liabilities were at face value or par. Federal government data were totals for U.S. government, federally sponsored credit agencies, and the Monetary Authority. Assets held in federal employee and railroad retirement funds appeared both as assets and equal offsetting liabilities. The state and local government data excluded employee retirement funds. No attempt was made to estimate and include retirement liabilities or associated future receipts at either the federal or state and local levels.

Conversion of values from par to market for financial assets and liabilities was accomplished as follows:

Currency, demand and time deposits, other loans, taxes receivable, and miscellaneous assets and liabilities. These items were taken at face value as indicated in the Flow of Funds accounts.

Gold. Prior to 1968, gold was taken at the official price of $35 per ounce. From 1968 through 1980, market values for the last trading day of the month were taken from the *Report of the Director of the Mint*, 1968, 1969,...1980, "Daily market prices of industrial gold in the United States, quoted by Engelhard Minerals and Chemical Corporation, Murray Hill, New Jersey."

U.S. government securities. Indices were taken or derived from published and unpublished data prepared by Cox and Hirschhorn. We applied the index for total U.S. government debt to Treasury issues, the index for U.S. Treasury notes to savings bonds and a weighted average of indices for Treasury bonds and notes to federal credit agency issues.

State and local obligations. For short-term obligations, that is, those of less than one year, we used the Cox and Hirschhorn index for U.S. Treasury bills. For other state and local bonds, we constructed an index by assuming that bonds have maturities of ten years, biannual interest payments, and coupon rates equal to the Standard and Poor's annual yields on long-term state and local bonds in the years they were issued. The volumes of new bond issues were taken from *Business Statistics* and from the *Federal Reserve Bulletin*. Market values were then calculated as the sum of interest and principal commitments, discounted by the Standard and Poor's end-of-year yield on state and local bonds, of the bonds issued over the ten most recent years.

Mortgages. We used a similar method in constructing a market-to-par index for mortgages, in this case assuming a 15-year life and discounting by the Federal Home-Loan Bank Board new home mortgage yield.

Corporate and foreign bonds (held by state and local governments). We used the New York Stock Exchange bond index for all listed U.S. companies, as found in the New York Stock Exchange *Fact Book*.

For net revaluations, we generally followed the formulation and procedures described in Eisner (1980). Thus,

$$\text{(A1)}\qquad NR_t = \left(K_t - \frac{P_{t,end}}{P_{t-1,end}} K_{t-1} - \frac{P_{t,end}}{P_t} IN_t\right) \frac{P_t}{P_{t,end}},$$

where NR_t = net revaluations of the year t, K_t = net value of stock at the end of the year t, P_t = the *GNP* implicit price deflator for the year t, $P_{t,end} = (P_{t,IV} + P_{t+1,I})/2$, with $P_{t,IV}$ and $P_{t+1,I}$ denoting *GNP* implicit price deflators for the fourth quarter of the year t and the first quarter of the year $t+1$, respectively, and IN_t = net investment or net flows in the year t.

For all financial assets other than gold and corporate equities, net investment in year t is the difference in the par value of stocks between years t and $t-1$.

---

[20] BEA Capital Stock Tape (1980). Data for 1980 were obtained by telephone from Jerry Silverstein.

[21] Flow of Funds Accounts, Division of Research and Statistics, Board of Governors of the Federal Reserve System, Flow of Funds data tape (September 1981).

Net investment in corporate equities is the net flow in the Flow of Funds accounts. Net investment in gold is equal to the change in the physical stock of gold, multiplied by the average market price.

For conversions of budget surplus or deficit to 1972 dollars, we used the *GNP* implicit price deflator. For gross public debt and net debt, we used our estimate of the end-of-year deflator, the mean of the fourth-quarter deflator and the deflator for the first quarter of the subsequent year.

The official high-employment budget surplus data were taken from de Leeuw and Holloway.

Interest rate effects equal the total change in par-to-market revaluations:

$$\text{(A2)}\quad IRE_t = \sum_{i=1}^{4} \left[(1 - MP_{i,t}) K_{i,t} - (1 - MP_{i,t-1}) K_{i,t-1} \right],$$

where $K_{i,t}$ is the par value of the ith financial instrument in year t, expressed as a positive quantity if it is a liability and a negative quantity if it is an asset, and $MP_{i,t}$, $i = 1$ to 4, are the market-to-par indices for the four financial instruments—savings bonds, mortgages, net holdings of agency issues, and other Treasury issues.

Price effects, expressed in current prices, relate to changes, due to inflation, in the real value of existing net debt exclusive of gold. They are calculated as the net revaluations described in (A1) above, but net of the interest rate effects defined in (A2), where K_t is the market value, exclusive of gold, of federal net debt at the end of year t, and IN_t is the market value of the change in net debt, exclusive of gold, in the year t.

REFERENCES

Boskin, Michael J., "Federal Government Deficits: Some Myths and Realities," *American Economic Review Proceedings*, May 1982, *72*, 296–303.

Cox, Michael and Hirschhorn, Eric, "The Market Value of U.S. Government Debt: Monthly, 1942–1980," *Journal of Monetary Economics*, March 1983, *11*, 261–72.

de Leeuw, Frank and Holloway, Thomas M., "The High-Employment Budget: Revised Estimates and Automatic Inflation Effects," *Survey of Current Business*, April 1982, *62*, 21–33.

Eisner, Robert, "Capital Gains and Income: Real Changes in the Value of Capital in the United States, 1946–1977," in Dan Usher, ed., *The Measurement of Capital*, Chicago: University of Chicago Press, 1980, 175–346.

______, "Social Security, Saving and Investment," *Journal of Macroeconomics*, Winter 1983, *5*, 1–19.

Fellner, William, "The High-Employment Budget and Potential Output: A Critique," *Survey of Current Business*, November 1982, *62*, 26–33.

Goldsmith, Raymond W., *Institutional Investors and Corporate Stock—A Background Study*, New York: National Bureau of Economic Research, 1973.

Horrigan, Brian R. and Protopapadakis, Aris A., "Federal Deficits: A Faulty Gauge of Government's Impact on Financial Markets," *Business Review*, Federal Reserve Bank of Philadelphia, March/April 1982, 3–16.

Leimer, Dean R. and Lesnoy, Selig D., "Social Security and Private Saving: New Time-Series Evidence," *Journal of Political Economy*, June 1982, *90*, 606–29.

Penner, Rudolph G., "How Much is Owed by the Federal Government?," *Carnegie-Rochester Conference Series on Public Policy: Monetary Regimes and Protectionism*, Spring 1982, *16*, 233–56.

Ruggles, Richard and Ruggles, Nancy D., "Integrated Economic Accounts for the United States, 1947–1980," *Survey of Current Business*, May 1982, *62*, 1–53.

Seater, John J., "The Market Value of Outstanding Government Debt, 1919–1975," *Journal of Monetary Economics*, July 1981, *8*, 85–101.

Congressional Budget Office, *The Outlook for Economic Recovery*, Washington: USGPO, February 1983.

U.S. Council of Economic Advisers, *Economic Indicators*, May 1983, Washington: USGPO, 1983.

______, *Economic Report of the President*, Washington: USGPO, February 1982.

Federal Reserve Bulletin, selected issues, 1979–82.

New York Stock Exchange, *Fact Book*, 1951...1980.

The Wall Street Journal, June 22, 1983.

U.S. Department of Commerce, Bureau of Economic Analysis, *The National Income and Product Accounts of the United States, 1929–74*, *Statistical Tables*, Washington: USGPO, 1976.

______, *Business Statistics*, Washington: USGPO, 1979.

______, BEA Control Stock Tape, BEA-N1W-80-002, Washington, 1980.

U.S. Department of the Treasury, Bureau of Government Financial Operations, *Statement of Liabilities and Other Financial Commitments of the United States Government as of September 30, 1982.*

______, Bureau of the Mint, *Report of the Director of the Mint*, 1968...1980.

Wharton Econometric Forecasting Associates, *Wharton Long-Term Forecast*, *April 1983*, Philadelphia: 1983.

Coping with Complexity: The Adaptive Value of Changing Utility

By MICHAEL D. COHEN AND ROBERT AXELROD*

In everyday life we frequently act on the basis of beliefs that are incomplete or, to some degree, false. Yet academic research on human choice behavior pays surprisingly little attention to the difficulties facing a decision maker whose current model of the world usually has a substantial probability of being misspecified. Most research can be viewed as providing potential elements of an improved future system of beliefs, one that will better match the world. But that very fact implies that today's decisions and those made earlier may well have been based on wrong beliefs. How should a decision maker behave under such conditions? When it is too costly to finance adequate investigation or to postpone decisions while awaiting its results, how might a reasonable, but not omniscient, person proceed using beliefs about the world that may be partially incorrect?

Existing literature gives only limited help on this question. One major research tradition deals with the updating of beliefs in view of new experience. This tradition includes both econometrics and Bayesian inference. These techniques can be very helpful in updating beliefs where new information can be incorporated into a model that is correctly specified. But they offer less help when the "improved" parameter estimates are being obtained within an incorrect model of how things work. Among the many difficulties that can occur with a misspecified model, one of the most trying is the possibility that such a model can make correct predictions over some range of policies. The apparent confirmation in this "misspecification trap" can prevent the decision maker from pursuing an improved model that would reveal superior alternatives remote from current policy.[1]

Our proposal for a response to the problems posed by wrong beliefs is a decision process that incorporates a controlled form of preference change. It sounds paradoxical to say that individuals may perform better as a result of moving away from their current goals. But we will show that a properly structured adaptive utility process can indeed help people make better decisions.

I. Misspecification and Preference Change

This paper offers a model which demonstrates that preference change can provide a course of action that is adaptive in the presence of misspecification. By "adaptive," we mean that such a process can improve performance. The preference change is driven by experience, but is not necessarily conscious.

Our model bears on both normative and descriptive issues. On the normative side, it

---

*Associate Professor and Professor of Political Science and Public Policy, respectively, Institute of Public Policy Studies, 1516 Rackham Building, University of Michigan, Ann Arbor, MI 48109.

We thank Joel Cohen, Jon Bender, John Ferejohn, Aanund Hylland, Lewis Kornhauser, James March, Robert A. Pollak, Richard Quandt, and John Scholz for their helpful suggestions. We are especially grateful to Carl Simon who generously devoted many hours to refining an earlier version of our model. Earlier versions of this paper were presented at the Conference on the Multiple Self, Maison des Sciences de l'Homme, Paris, and the Carnegie-Mellon Symposium of Information Processing in Organizations. This work was supported by a Sloan Foundation grant to The University of Michigan Institute of Public Policy Studies, the Center for Advanced Study in the Behavioral Sciences, and NSF grant SOC-8023556 to Axelrod, and NSF grant SES-7726026 to Cohen.

[1] The other major research tradition dealing with the problem of inaccurate beliefs, investigates how people of limited rationality can still cope with their environment (for example, James March and Herbert Simon, 1958; Richard Cyert and March, 1963). Concepts such as satisficing and dynamic aspiration levels do suggest that decision mechanisms as well as beliefs are adaptable. Yet even in the research tradition of limited rationality, the underlying preferences are taken as given.

provides an approach to the problem of improving your performance when you don't completely understand what you are doing. On the descriptive side, the approach seeks to account for some of the ways in which preferences do in fact change as a function of experience. We seek not only to model adaptive preference change, but to do so using assumptions that are behaviorally plausible.

Economists interested in how tastes change have taken two basic approaches. The sophisticated behavior approach of Robert Strotz (1955–56) assumes that individuals know that their present choices influence their future preferences. This gives rise to a variety of problems of consistency, existence and stability of plans and choices over time (Carl Christian von Weizsacker, 1971; Bazalel Peleg and Menahem Yaari, 1973; Robert Pollak, 1968; Peter Hammond, 1976; and Gordon Winston, 1980). Besides these technical problems, there is also the practical question of whether a decision maker is likely to know much about how a current choice would affect future preference.

In contrast, the naive behavior approach assumes that the decision maker does not know how current choice would influence his or her future preferences. This approach has been used in empirical demand analysis. The work has explored four ways in which tastes change over time: habit formation, interdependence between people, advertising, and price signals. (For a review, see Pollak, 1978.) For our purposes, the most relevant strand of this literature is habit formation, the idea that future preferences depend directly upon prior choices. Habit formation is typically operationalized by providing a functional relationship between the amount of some current (or past) activity and the parameters of the next period's utility function.[2] This captures the concept of a habit developing as a function of action, but it leaves out the fact that beliefs may also have influenced the shaping of preferences.

The model of preference change which we propose remedies this omission by explicitly introducing the interaction between beliefs and actions in the shaping of future preferences. An action taken on the basis of beliefs can yield surprise. We define "surprise" as the difference between the utility experienced as the result of an action and the utility expected to result from that action. We then model an individual as coming to like the things that yield pleasant surprises and coming to dislike the things that yield unpleasant surprises.

The model forges a connection between the literature on changes in preferences and the literature on rules of thumb and bounded rationality. It does this by viewing preference change as being driven by the surprises that will inevitably occur if the world is too complex and dynamic for the decision maker to develop a correctly specified model of the environment. Because of such complexity and the resulting misspecification of belief, we do not want to assume that the decision maker can anticipate his or her own future preference changes. Instead, we are studying the situation in which utility change represents a largely unconscious adaptation to an environment which is not completely understood.

To illustrate our basic approach, consider the game of chess. With a correctly specified model, one could choose the best play at every turn, but the explosive combinatorics of chess have so far (and for the foreseeable future) prohibited the development of a correct model. Play must be heuristic, based on principles known to be imperfect. The goal, of course, remains the capture of the opponent's king. But at the beginning and the middle of the game, the player cannot see just how to accomplish this goal. So the player pursues other goals as well, hoping that will lead toward improved performance even on the criterion of winning the game. A chess player is typically taught to evaluate features of board positions in particular ways. For example, a rook is valued at five points, a knight at three points, and a pawn at one point. These valuations lead directly to useful policy advice. For instance, they suggest that a player should be willing to give up a knight to capture a rook, but should not be willing to make the reverse trade. These values for the pieces are not specified in the rules of the game. But neither are they arbitrary. They are the result of centuries of experience. The valuation of the pieces and of other aspects of board positions by successful players indicates that one can do better by *not* concentrating exclusively on the capture of the king, but by learning also to pursue goals that eventually make the ultimate goal more accessible.

---

[2] George Stigler and Gary Becker (1977) have proposed a human capital theory approach to account for certain self-reinforcing patterns of behavior. But their theory accounts for the changes by alterations in "technology" (for example, the ability to make finer discriminations), rather than modifications in the underlying utility itself. This approach has been criticized for being both a doubtful ideology and having no practical relevance for normative prescriptions of choice (March, 1978, p. 597). The issue has also been tackled by Cyert and Morris DeGroot (1975; 1980), but their approach is to study how experience alters a person's *knowledge* of what he or she likes, rather than how it alters underlying utility.

Our model is actually a direct descendant of an artificial intelligence program by A. L. Samuel (1959) for playing checkers. The success of the Samuel Checker Player could be evaluated by an outsider using a simple performance measure: whether or not it wins games. The program itself was not given the set of values to be used in the evaluation of board positions. Instead it learned for itself what values to follow, beyond the basic one of piece advantage. The program's learning process was driven by surprise. When things were going surprisingly well (or badly), it would note what was correlated with surprise, and would adjust the values of its various goals accordingly. To take the chess analogy again, suppose a person valued a rook and a knight equally. Then when the player gave up a rook to capture a knight, the player might notice that a few moves later things were going surprisingly badly. If unpleasant surprises frequently occurred after giving up a rook to get a knight, the learning process would gradually raise the relative value imputed to rooks. Thus surprises can drive an adaptive change in values. This is just how Samuel's checker playing program learned to play good checkers.[3] In fact, the program was able to defeat a former state champion (Edward Feigenbaum and Julian Feldman, 1963, pp. 103–104).

This paper seeks to generalize the principles that make the Samuel checker playing program so successful. To demonstrate that the principles are applicable beyond checkers, we want to show their successful operation in another task domain. To maximize the clarity of the principles, we have used as simple a setting as we can without making the choice trivial.

II. A Dynamic Model of Preference Change

An extended example will illustrate both the nature of the problem we have in mind and the application of the principles that we believe to be useful in responding to the problem. We introduce some notation in order to make the inferences in the presentation precise, but will make a number of simplifying assumptions so that the mathematical details do not obscure the issues at hand.

Consider a factory manager with a fixed number of labor hours available for the upcoming period. The manager wants to maximize output and has to allocate labor hours between production and maintenance.

The manager knows that too much labor devoted to production would lead to suboptimal output because of inadequate maintenance. The manager also knows that too little labor devoted to production will lead to suboptimal output. Therefore the manager's problem is to choose the level of labor devoted to production, x, which maximizes output, y. We will assume that the relationship between x and y is *believed* to be of the following form:

$$\hat{y}_t = -x_t^2 + \hat{b}_{t-1}x_t, \tag{1}$$

where $\hat{y}_t$ is the expected output, and $\hat{b}_{t-1}$ is a parameter that can be estimated from the previous choice of x and the observed output in the previous time period.[4] This functional form is typical of a broad class of problems where increasing the level of activity is beneficial at first, but then becomes detrimental.

---

[3] Chess programs typically have fixed evaluation functions, and hence do not learn. The learning in computer chess takes place in the programmer's mind and is embodied in the next version of the program.

[4] For example, this relationship could have come from an analysis of labor productivity. The manager

With a choice of x_t, and an observed level of output, y_t, the manager can estimate the unknown parameter in equation (1) by solving for $\hat{b}_t$. This gives (for x_t not equal to zero):

$$(2) \qquad \hat{b}_t = y_t/x_t + x_t.$$

With this estimate of the unknown parameter, the manager can select the next policy choice to maximize expected output. A little calculus shows that the x_{t+1} which maximizes $\hat{y}_{t+1}$ is

$$(3) \qquad x_{t+1} = \hat{b}_t/2.$$

Now we are in a position to introduce the problem that we want to study, namely misspecification.[5] Equation (1) gives the *believed* production relationship. Our interest is in the case in which this belief is inaccurate. So we will suppose that there is an unknown source of lost output in the factory due to pilferage. We will denote the number of units being lost each period as c ($c < 0$) and make the *true* output relationship

$$(4) \qquad y_t = -x_t^2 + bx_t + c.$$

We will assume that this simple misspecification will not be discovered by the manager. The reason for this assumption is to allow us to embody the important principle that the world is so complex that there is always *something* that beliefs do not model adequately. We could easily endow the manager with a more powerful belief system capable of discovering the missing variable, and then increase the complexity of the reality in order to keep the belief system inadequate. Beliefs would still fall short of reality, but the mismatch would occur at a much higher level of complexity. This might add an appearance of realism, but at the cost of burdening our exposition with details that would obscure the structure of the argument without strengthening its fundamental logic.

The manager's method for choosing an optimal level of x is no longer ideal in light of this misspecification, but it will be used since the loss of c units each period is not known. The level of x chosen according to equation (3) will therefore no longer be optimal. If the manager sticks to the choice procedure, there will be a brief discrepancy between expected output and observed output. After that, the policy choice will settle down to

$$(5) \qquad x_s^* = \left(b + (b^2 + 8c)^{1/2}\right)/4.$$

We will refer to the model of the manager's decision making that we have just sketched as the *standard model* so that we may contrast it with a dynamic model that we will develop below. The allocation x_s^* will be called the *standard result*.[6] It does not depend on what initial policy was tried to provide the first estimate $\hat{b}_1$, and it is suboptimal since the standard result will be less than $b/2$ whenever $c < 0$. Furthermore, the standard result has the devilish property that it produces a misspecification trap: output experience will exactly match expectations based on incorrect beliefs.

Figure 1 shows the geometry of the situation the unfortunate manager is in. The two

---

might believe that $\hat{Y}_t = kp_t x_t$, where k is a constant of proportionality, $p_t = q + r(L - x_t)$ where q is the (unknown) minimum productivity of a work hour that would occur if all the available labor were allocated to production, r is the (known) rate at which a labor hour invested in maintenance improves productivity, and L is the (known) total labor available. Then if we rescale the output by letting $\hat{y}_t = \hat{Y}_t/kr$, we have equation (1) where $\hat{b} = L + q/r$. Note that $\hat{b}$ is unknown because q is unknown.

[5] The mechanisms for updating beliefs and choosing new policy incorporated in (2) and (3) are perhaps a bit more precise than one might expect from real decision makers. Our decision to make new beliefs perfectly consistent with the most recent experience, and new policy optimal with respect to those beliefs, rests on a desire to provide the most stringent test of the contribution of dynamic preferences to decision-making quality. With these forms for (2) and (3), we can be more confident that the performance improvement obtained is not merely a correction for faulty methods of updating beliefs or choosing policy.

[6] Equation (5) is derived by noting that at stability equations (2), (3), and (4) imply that $x = (b + c/x)/2$. The right-hand side of (5) is just the largest root of this quadratic. To avoid an imaginary term we need $b^2 + 8c$ to be nonnegative. We restrict our discussion to cases satisfying this constraint.

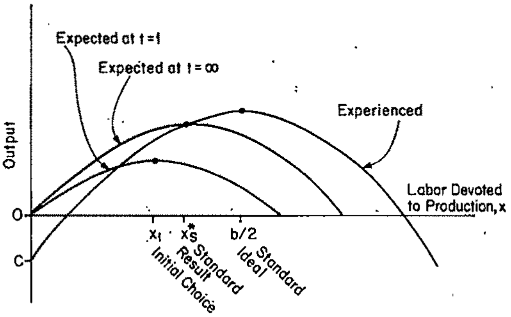


FIGURE 1. A MISSPECIFICATION PROBLEM

parabolas passing through the origin represent what is believed about the system initially and when the adjustment of beliefs to experience has settled down. The choice that would maximize output, given the final beliefs, is x_s^*. The true output relationship is shown by the other parabola. At the standard result it has the same value as the believed parabola and the incorrect model is therefore confirmed by experience.

To develop the dynamic model, we need an explicit expression for the manager's revealed utility. We will suppose it is the simple sum of output, about which the manager has explicit beliefs, and an additional factor that we call the *intrinsic utility* of x, the extent to which the manager regards production labor as a good (or bad) thing in itself, as a final value rather than as an instrument in the production of valued output. Intrinsic utility need not be conscious. When the manager places positive intrinsic value on x, this amounts, of course, to reducing the relative value of output in determining the choice of x. As we will show, however, the properly controlled evolution of the intrinsic value for x can yield an *increase* in output—the manager does better in output terms by caring less about it. The *expected utility* which the manager's choice will be maximizing is

$$(6) \qquad \hat{U}_t = \hat{y}_t + w_t x_t,$$

where w_t is the intrinsic utility of a unit of x. The manager's satisfaction with observed results, *experienced utility*, is output y_t plus intrinsic utility $w_t x_t$, or

$$(7) \qquad U_t = -x_t^2 + bx_t + c + w_t x_t.$$

For $w_t = 0$, (6) and (7) correspond to the standard model. Expected utility is maximized at $(\hat{b}_t + w_t)/2$. When w is a function of experience, we have our dynamic model. The dynamic choice process is like the standard process just presented in every respect except that in the dynamic case intrinsic utility evolves as a function of the discrepancy between expected and experienced utility, a quantity that we will call *surprise* and label D_t.

$$(9) \qquad \text{Surprise} = D_t = U_t - \hat{U}_t = (b - \hat{b}_{t-1})x_t + c.$$

Our proposed model is in essence a kind of learning process through which the manager comes to ascribe additional value to the assignment of labor to production if such assignments are associated with pleasant surprises, and comes to assign less value to the level of x when it is associated with negative surprises.[7]

We must specify a functional form governing the change in w and there are many possibilities. The field is narrowed sharply, however, by two constraints. First, the change in w should be proportional to the magnitude and direction of recent change in x. This can be represented as $(x_t - x_{t-1})/|x_t|$. The divisor is necessary to scale the expression so that it will depend only on the relative magnitude of the change. Second, the change in w should be proportional to the (scaled) magnitude and direction of recent surprise. This is just D_t/x_t. There are several ways these two expressions could be combined. Our choice is the relatively conservative position that change in w should be large only when surprise and policy change are

---

[7] Our decision to make surprise a fundamental quantity in our model follows Samuel. It is supported by recent findings that surprise is a genuine psychophysical state with reliably measurable corresponding brainwave patterns (Connie Duncan-Johnson and Emanual Donchin, 1977). Moreover, this measurable surprise can apparently derive from both conscious and unconscious expectations in correspondence to our distinction between instrumental and intrinsic utility. In cognitive psychology, an approach in the same spirit has been made by George Mandler (1981), who argues that changes in value are driven by a form of cognitive discrepancy or "schema incongruity."

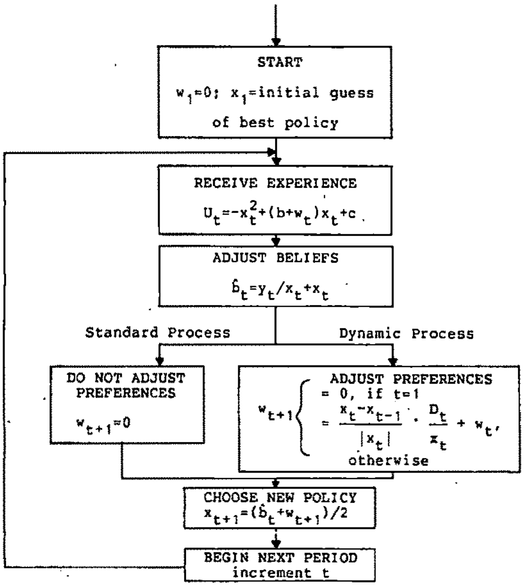

START
$w_1=0$; x_1=initial guess of best policy
RECEIVE EXPERIENCE
$U_t=-x_t^2+(b+w_t)x_t+c$
ADJUST BELIEFS
$\hat{b}_t=y_t/x_t+x_t$
Standard Process
Dynamic Process
DO NOT ADJUST PREFERENCES
$w_{t+1}=0$
ADJUST PREFERENCES
$w_{t+1}\begin{cases}=0, \text{ if } t=1\\ =\frac{x_t-x_{t-1}}{|x_t|}\cdot\frac{D_t}{x_t}+w_t, \text{ otherwise}\end{cases}$
CHOOSE NEW POLICY
$x_{t+1}=(\hat{b}_t+w_{t+1})/2$
BEGIN NEXT PERIOD
increment t

FIGURE 2. FLOWCHART FOR STANDARD AND DYNAMIC PROCESSES

both large.[8] This gives us an overall expression for the change in w as the product of these two expressions:

$$w_{t+1} - w_t = \frac{x_t - x_{t-1}}{|x_t|} \cdot \frac{D_t}{x_t}. \tag{10}$$

This functional form captures three important properties that we believe to be essential for the satisfactory operation of the dynamic process: surprise as the main driver of the adaptation of preferences; drag on the adaptation process to keep it from runaway pathologies; and implicit context sensitivity, which keys the changes in preferences to the conditions associated with the surprise. In our simple setting, these are appropriate analogues of the procedures employed in the Samuel Checker Player.

The flow chart of Figure 2 presents the complete cycle that is followed by our dynamic choice process, and includes the standard choice process as a special case in which w_t is always zero. We have not been able to obtain a closed-form expression for the *dynamic result*, the value of x (denoted x_d^*) that is obtained from running the dynamic process under various initial conditions. It is easy to compute this value, however, for a great variety of initial values, x_1, and for many levels of c, the specification error. We now turn to the results of such computations.

III. Assessment of Performance

Thus far we have delineated a pair of processes that a decision maker could use when seeking to maximize utility in a world that is not completely understood. The dynamic model and the standard model both have a structure of beliefs about the world built from the same (mis)specification of reality. Both have a system for updating those beliefs as new experience comes in. Both choose new policies so as to maximize expected utility given current beliefs. Both fail to choose optimal policies because of misspecification. The utility functions being maximized are, however, not identical, since the dynamic model has a method of changing its utility function that is inoperative in the standard model. This is, in fact, the only difference between the two models. It does mean, however, that the objective function being pursued by the dynamic model will, in general, be different from that being pursued by the standard model.

This last fact raises the question of how we are to assess performance. The standard model may be regarded as embodying the process that would occur if no preference change were allowed. The dynamic process begins with the same utility function as that used throughout the standard process. However, the dynamic model becomes, in effect, a different person. If we use the standard model's utility function to assess the quality of the policy chosen by the dynamic model, or if we use the dynamic function to assess standard model policy choices, we will be making an intertemporal comparison of utilities. The difficulties this raises are very similar to those present in the more common case of interpersonal comparison. We resolve the problem as has been done so often in the more familiar setting, by invoking the standard proposed by Pareto: if one outcome is preferred to another on *both* utility functions, we will call it the better of the two. This criterion is restrictive in many ways. We do not mean to propose that normative advice that fails to satisfy the criterion (i.e., that leads to outcomes not preferred on both the current and the future utility functions) is necessarily bad advice. But we do think that advice which leads to preferred outcomes on both utility measures will definitely be more acceptable to the potential taker—who, after all, has the standard utility function at the moment and will have the dynamic function later.[9] Therefore, we wish to determine the conditions under which the dynamic model is better than the standard model in this Pareto sense.

---

[8] Since two policy choices and observed outcomes are needed to scale the surprise, the adjustment in w does not start until $t = 2$.

To this end we introduce some additional definitions. The *standard ideal* is the policy at which the utility function of the standard model would be maximized. In our model, its value is $b/2$. The standard ideal might also be characterized as the "instrumental ideal," since it is the level of the activity x that maximizes utility when x is only an instrument and has no value of its own for the decision maker. Except when $c=0$, the standard ideal differs from the standard result of (5), the policy normally reached by the standard model.

The *utility of the standard result* (U_s^*) is the value of equation (7) with $w=0$ and $x=x_s^*$. A comparable quantity is the *utility of the dynamic result* (U_d^*) obtained from equation (7) with $w=0$ and $x=x_d^*$.

When the inequality

$$(11) \qquad U_d^*/U_s^*>1$$

is satisfied, the net output at the dynamic result is greater than at the standard result. Since the standard utility function values only output, we have a situation in which the dynamic process has come closer to the standard ideal than has the standard process. In this case, the standard utility function will have a higher value at the dynamic result than at the standard result, establishing the difficult part of the Pareto criterion. The other part, as one might expect for a rationalizing system, follows immediately, since a result that is better on the standard utility function is better on the dynamic utility function by an addition of the term for intrinsic utility, $w_t x_t$. (In the calculations we have performed, $w_t x_t$ is always positive if U_d^* is at least as large as U_s^*.)

Thus we have isolated the crucial measure to be used in assessing the performance of the dynamic model. When $U_d^*>U_s^*$, the dynamic result may be said to be better than the standard result in the Pareto sense. In terms of the chess example we have employed, this is equivalent to observing that a player who comes to care about objectives other than capturing the king may not only be happy in terms of the new objectives being pursued, but also may end up capturing the king more often—although the king's capture is no longer the sole objective.

We are now ready to turn to the results of the computations. Figure 3 is a topographic map, showing the performance of the dynamic model relative to the performance of the standard model. The results are given for particular values of the misspecification c, and the initial choice x_1. The computations were made with x_1 from just above zero through a starting value four times the standard ideal, and with c from just below zero to $-b^2/8$, the minimum for which the standard process converges. The unhatched portion of the surface corresponds to the set of parameter values for which the utility ratio is greater than one. These are the points where the dynamic process achieves more output than the standard process. It is the overwhelming majority of the surface.

The possible values of the initial belief x_1 should not be imagined to be equiprobable, however. For that reason we have shown what might be considered the most probable area of the figure by adding two vertical dotted lines at one-half and twice the standard ideal. A manager beginning a sequence of choices with the benefit of reasonable

---

[9] It is interesting to note that a common form of preference change—imitating those one already likes—also satisfies this Pareto logic. The imitator already likes the behavior of the imitated and can expect to like it more as the movement continues toward the imitated's values.

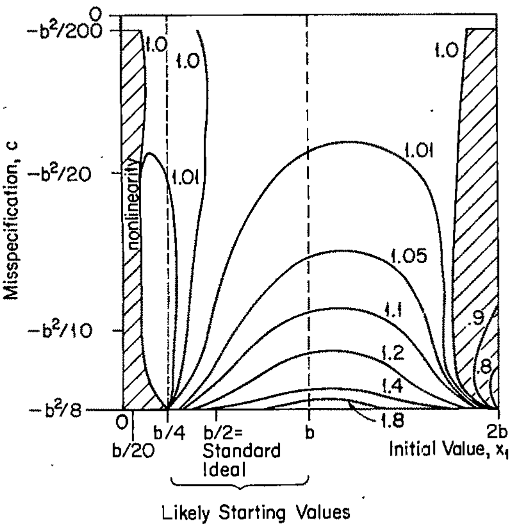


FIGURE 3. RATIO OF PERFORMANCE: DYNAMIC/STANDARD

prior beliefs derived from imitation or analysis would be quite likely to start at a point in this central area. In this region of plausible initial beliefs, the dynamic process completely dominates the standard process. Thus, with any initial belief that differs from the standard ideal by even as much as a factor of two, the dynamic process does better than the standard process.

When the pilferage rate c is high, the results are near the bottom of the figure, where the dynamic model does up to twice as well as the standard. For values of x_1 very close to zero, there are sharp nonlinearities in the behavior of the dynamic model, but they are confined to a small region well distant from plausible starting values for the dynamic process. Some points here do satisfy inequality (11), but others nearby do not.

Figure 3 verifies that the final policy reached from likely starting conditions is always better in the dynamic case. Our computations actually give the stronger result that from all likely initial conditions, the cumulative output of the dynamic model exceeds that of the standard model at every period. The dynamic model is better not only when stability is reached, but throughout the entire process of adaptation.

The consistent dominance of the dynamic result over the standard result displayed in the portion of the figure between the dotted lines demonstrates an important point. Under a wide range of initial conditions, a decision maker in this model would be well-advised to let his or her values drift away from their current configuration in a properly controlled adaptation to experience. Over a wide range of parameter values, a process that is essentially a controlled form of rationalization leads to results that will be preferable both as measured by the current utility function and as measured by the utility function at the conclusion of the dynamic process.

Once the dominance of the dynamic model over the static model has been established, a further question asserts itself. How robust is this superior performance of the dynamic model in misspecified environments? There is an enormous number of model variations that might be run to investigate robustness, but one line of such variations seems particularly important. These are versions of the models that test whether the strong results obtained depend on the particular form of misspecification used. The mismatch between reality (equation (4)) and belief (equation (1)) in the case that was studied was the omission of a constant ($c < 0$) from the belief system.

A variation on the experiment that suggests itself immediately is the symmetric misspecification in which the omitted term (c) would be greater than zero. We have repeated our calculations with all the variables covering the same absolute ranges as before and c positive. The results are essentially the same. None of our qualitative conclusions are affected. In particular, it is worth noting that this demonstrates the full symmetry of the dynamic model's performance: its adaptive contribution occurs whether the standard result and initial policy are above or below the best policy, and whether the ultimate value of w is positive or negative.

To assess robustness further, we subjected the dynamic and standard models to more radical misspecifications in three additional experiments. In each case we substituted a new reality for equation (4) of our models. All other aspects remained unchanged. The beliefs about the production relationship continued the assumption that it was a parabola passing through the origin. The method of updating beliefs based on experience operated as before. The method of choosing a new policy expected to be optimal in a quadratic world was retained. In the dynamic model, the preference change process driven by surprise was identical. All that was different was the nature and magnitude of the mismatch between belief and underlying reality.

In the three robustness experiments, the new realities were

$$(4') \quad y_t = -|x_t - b| + c, \quad \text{(absolute value)};$$

$$(4'') \quad y_t = -\sin(cx_t + b), \quad \text{(sine)};$$

$$(4''') \quad y_t = x_t^3 - x_t^2 + bx_t + c, \quad \text{(cubic)}.$$

All three were scaled with respect to the variables x_1 (initial policy) and c (the omitted parameter) so as to be comparable to the original case. All three have only a single maximum in the range studied, but the curvatures in the range are all quite different from the original reality function and therefore from the parabola incorporated in the belief systems of both models. As a result, the models behaved quite differently in the three additional cases. However, the fundamental qualitative conclusion of the original case was sustained in all three of the new studies: over the same wide range of variation in initial policy (x_1) and in magnitude of the omitted parameter (c), the dynamic model outperformed the standard model. Its cumulative performance on both measures was better at every step along the way.

Thus the robust superior performance of the dynamic model demonstrates that the simple principles underlying the success of the Samuel Checker Player can be transferred with powerful effect to other task domains.

IV. A Management Example

Consider a newly hired middle manager who begins with a concern only to win the approval of his or her superiors. Proceeding from that orientation, such a manager might initially join them in opposing an expensive package of fringe benefits for the workers in the manager's unit. Over time, however, the operation of the dynamic process we described might well produce a growing sensitivity to winning the approval of subordinates if there is some regular but poorly understood relationship between subordinate approval and performance by the unit that satisfies top management. In turn, such a development can find the manager coming eventually to support the fringe benefit package that the workers want, getting better performance out of the unit and therefore winning more net approval of superiors. Such a result is consistent with a pattern frequently observed in organizations. Consider, for example, Peter Blau's report (1956) of a study by Daniel Katz et al. (1950): "...superiors who were primarily concerned with maintaining a high level of production, interestingly enough, were less successful in meeting this goal than those supervisors who were more interested in the welfare of their subordinates than in sheer production; in the latter case, productivity was generally higher" (p. 70).

All this can occur in the dynamic process without the manager ever having a full understanding of the causal connections producing the effects. It seems to us consistent with the observation that managers (and other regulators) are sometimes "captured" by the interests of those they manage. It also suggests that in a complex environment, a suitably controlled capture process may have some genuine virtues.

V. The Interpretations of Our Model

This example can be used to illustrate two different interpretations of our model. The conventional interpretation is that the manager only cares about production, and treats the welfare of the subordinates as a means to that end. Under this interpretation, the manager's utility is simply a function of production. The conventional interpretation would then regard the mechanism for using surprise as a "rule of thumb" which shapes a pseudo-utility function.[10] The interpretation that we prefer drops the concept of a pseudo-utility function, and regards the manager as having undergone a genuine change in preferences. Certainly, this would be the natural interpretation from the perspective of revealed preference theory: if the manager is observed to be willing to give up some expected production to get some expected employee welfare, then this choice can be taken as reflecting a utility function which includes both production and welfare.

An individual need not realize that he or she is coming to like things that are associated with pleasant surprises—but the resulting change in the utility function can still be adaptive. As long as there is misspecification, there is need for more than the usual Bayesian approach by which new information is incorporated into the decision process through the revision of beliefs (Sanford Grossman, Richard Kihlstrom, and Leonard Mirman, 1977; and John Hey, 1981). Under our preferred interpretation, it would be reasonable to say that new information is used not only to update beliefs about how choices map into outcomes, but that it is also used to modify the very utility function that is being maximized. Indeed, we regard the demonstrated potential for adaptive utility change as the fundamental point of our model.

VI. Learning to Adapt in a Strategic Interaction

The theory of adaptive utility change has so far been discussed in the context of a single decision maker in a passive environment. The theory can also help to explain behavior in a setting of mutual interaction. Consider for example the iterated Prisoner's Dilemma in which a player chooses in each turn whether to cooperate or defect. A simple strategy in the game is Tit for Tat: cooperate on the first move and then do what the other player did on the previous move. This strategy has been shown by Stuart Oskamp (1971) and Warner Wilson (1971) to be highly successful when playing directly with people. It has also been shown by Axelrod (1980; 1981) to be highly successful when playing with a wide variety of more or less sophisticated decision rules. However, even a naive player without a clear knowledge of the other's strategy can realize that there is instrumental utility to defecting. After all, the logic of Prisoner's Dilemma guarantees that on a given move the payoff from defecting will always be greater than the payoff from cooperating, no matter what the other player chooses. But there is a pleasant surprise in store for the naive player who experiments while interacting with Tit for Tat. In the move after the naive player cooperates, the player gets a higher payoff than in the move after a defection is chosen. According to the model of adaptive utility developed here, this pleasant surprise will actually cause the naive player to place some intrinsic utility on the choice of cooperation after the other player cooperates. When enough intrinsic utility has been assigned to this conditional cooperation, the player will overcome the instrumental utility of defection, and come to cooperate when appropriate.

The trick in the Prisoner's Dilemma is to cooperate only with those who will reciprocate. This is already built into Tit for Tat. It is a trick that someone employing an adaptive utility process can learn from scratch.

VII. Conclusions

Beliefs are virtually always misspecified to a greater or lesser extent. Our results show that when beliefs are misspecified, controlled preference change can actually be adaptive. By allowing pleasant and unpleasant surprises to guide changes in utility, a decision maker can actually achieve better performance on both the original and revised utility functions. This possibility has now been demonstrated not only in the complex task environment of checkers, but also in four variations of a simple model of a management task.

---

[10] Rules of thumb have been investigated by William Baumol and Richard Quandt (1964) in the context of optimally imperfect decisions, namely decisions where the marginal cost of additional information or calculation equal the marginal expected yield. In our model, the amount of information is fixed, but the problem remains of how best to use it.

The significant novelty in our approach is the way in which preferences change as a function of experience. We believe the success we have had rests substantially on incorporating—as did Samuel in the Checker Player—three principles of preference change: 1) surprise as the main driver of adaptation; 2) drag to retard excess value change; and 3) implicit context sensitivity.

The first of these is essentially the principle that the utility function adapts not to experience itself, but to the difference between experienced and expected utility. In effect, the adaptation of the function is to the errors of the cognitive system: today's intrinsic utilities are in part the result of yesterday's misunderstandings of the world. The second characteristic of our treatment of utility dynamics is the use of drag on the process of bringing utilities into alignment with experience. If this process is allowed to occur too rapidly, there is a serious risk of a "runaway" utility function that dictates extremely high or extremely low levels of the activities in question. We regard this model pathology as akin to the development of an addiction or phobia. While a too-rapid adjustment is to be avoided, so also is too much drag. At the extreme, excessive drag would leave a dynamic process nearly indistinguishable from the standard one, and the benefits available would have been forfeited to caution. In his stimulating 1978 paper, March observed that the a posteriori adjustment of preferences toward consistency with outcomes ("rationalization" in common parlance) might not be maladaptive in every case. He lamented the lack of a formal framework in which one could pursue the issue with precision. Our model provides the first formal treatment of the question, so far as we are aware.

The third characteristic of our treatment of value change is an effort to arrange the adaptive process so that it implicitly incorporates as much sensitivity to context as possible. In the dynamic model, this occurs because the sign of the change in intrinsic utility depends on whether or not recent changes in policy (and, implicitly, recent changes in intrinsic utility) have been positive or negative. In the Samuel Checker Player, a considerably more elaborate system accomplished a similar result.

An issue addressed in only its simplest form by our model is the attribution problem: to which of the many things going on now or recently should the system give the credit (or blame) for the surprise it has experienced? Some portion of this dilemma is resolved by whatever cognitive mechanisms are available for updating the system's beliefs. But it is unlikely that these mechanisms will be entirely adequate, so that discrepancies will probably continue to occur. The intrinsic values that rise or fall in such circumstances need to have a better than random chance of being genuinely related to the experienced discrepancies if the learning is to be anything but superstition.[11] Samuel was able to specify a method, based on feature correlations, which handled the attribution problem in the special case of checkers, but we are not certain what the correct generalization of his methods will be.[12]

The success of the three simple principles of preference change that we have followed is certainly encouraging. They led us, almost without a false step, to the model we have presented, they seem to be consistent with the procedures used in Samuel's remarkable checker program, and they are generally consistent with results of psychological research. We can, at this point, only claim success for the approach in the special case presented here. But this special case does serve as an existence proof that utility changes which are guided by surprise can be adaptive. If this principle turns out to be broadly applicable, it would have important consequences for fundamental questions in economics, political science, and organization theory.

---

[11] As John Anderson notes (1980), the most successful models of learning and cognition have postulated a fundamental capability of responding to features that are correlated in the environment. This process need not be conscious. Indeed, Anderson et al. (1979) have shown that subjects often cannot verbalize the bases of their learned (and correct) experimental responses. They sometimes even exhibit behaviorally powerful unconscious expectations which are directly at odds with their conscious models of the experimental situation (Daniel Kahneman and Amos Tversky, 1982).

[12] One possible solution may lie in employing an analogue of the adaptive process used by a pool of genes to become increasingly more fit in a complex environment. A promising effort to convert the main characteristics of this process to an heuristic algorithm is given by John Holland (1975). This algorithm has had some striking preliminary success in the heuristic exploration of arbitrary high dimensionality nonlinear functions.

REFERENCES

Anderson, John, *Cognitive Psychology and Its Implications*, San Francisco: Freeman, 1980.

______, **Kline, Paul and Beasley, Charles,** "A General Learning Theory and Schema Abstraction," in G. Bower, ed., *The Psychology of Learning and Motivation: Advances in Theory and Research*, Vol. 13, New York: Academic Press, 1979.

Axelrod, Robert, "More Effective Choice in the Prisoner's Dilemma," *Journal of Conflict Resolution*, September 1980, *24*, 379–403.

______, "The Emergence of Cooperation Among Egoists," *American Political Science Review*, June 1981, *75*, 306–18.

Baumol, William J. and Quandt, Richard E., "Rules of Thumb and Optimally Imperfect Decisions," *American Economic Review*, March 1964, *54*, 23–46.

Blau, Peter M., *Bureaucracy in Modern Society*, New York: Random House, 1956.

Cyert, Richard M. and DeGroot, Morris H., "Adaptive Utility," in R. H. Day and T. Groves, eds., *Adaptive Economic Models*, New York: Academic Press, 1975, 223–46.

______ **and** ______, "Learning Applied to Utility Functions," in Arnold Zellner, ed., *Bayesian Analysis in Econometrics and Statistics*, New York: North-Holland, 1980, 159–68.

______ **and March, James G.,** *A Behavioral Theory of the Firm*, Englewood Cliffs: Prentice-Hall, 1963.

Duncan-Johnson, Connie C. and Donchin, Emanual, "On Quantifying Surprise: The Variation of Event Related Potentials with Subjective Probability," *Psychophysiology*, September 1977, *14*, 456–67.

Elster, Jon, *Ulysses and the Sirens: Studies in Rationality and Irrationality*, Cambridge: Cambridge University Press, 1979.

Feigenbaum, Edward A. and Feldman, Julian, *Computers and Thought*, New York: McGraw-Hill, 1963.

Grossman, Sanford J., Kihlstrom, Richard E. and Mirman, Leonard J., "A Bayesian Approach to the Production of Information and Learning by Doing," *Review of Economic Studies*, October 1977, *44*, 533–47.

Hammond, Peter J., "Changing Tastes and Coherent Dynamic Choice," *Review of Economic Studies*, February 1976, *43*, 159–73.

Hey, John D., "Are Optimal Search Rules Reasonable? And Vice Versa?," *Journal of Economic Behavior and Organization*, March 1981, *2*, 47–70.

Holland, John H., *Adaptation in Natural and Artificial Systems*, Ann Arbor: University of Michigan Press, 1975.

Kahneman, Daniel and Tversky, Amos, "Variants of Uncertainty," in D. Kahneman et al., eds., *Judgment Under Uncertainty: Heuristics and Biases*, New York: Cambridge University Press, 1982.

Katz, Daniel, MacCoby, Nathan and Morse, Nancy C., "Productivity Supervision, and Morale in an Office Situation," Institute for Social Research, University of Michigan, 1950.

Mandler, George, "The Structure of Value: Accounting for Taste," Center for Human Information Processing Report 101, University of California-San Diego, May 1981.

March, James G., "Bounded Rationality, Ambiguity, and the Engineering of Choice," *Bell Journal of Economics*, Autumn 1978, *9*, 587–608.

______ **and Simon, Herbert A.,** *Organizations*, New York: Wiley & Sons, 1958.

Oskamp, Stuart, "Effects of Programmed Strategies on Cooperation in the Prisoner's Dilemma and Other Mixed-Motive Games," *Journal of Conflict Resolution*, June 1971, *15*, 225–59.

Peleg, Bazalel and Yaari, Menahem E., "On the Existence of a Consistent Course of Action when Tastes are Changing," *Review of Economic Studies*, January 1973, *40*, 391–401.

Pollak, Robert A., "Consistent Planning," *Review of Economic Studies*, April 1968, *35*, 201–08.

______, "Habit Formation and Long-Run Utility Formation," *Journal of Economic Theory*, October 1976, *13*, 272–97.

______, "Endogenous Tastes in Demand and Welfare Analysis," *American Economic Review Proceedings*, May 1978, *68*, 374–79.

Samuel, A. L., "Some Studies in Machine Learning Using the Game of Checkers," *IBM Journal of Research and Development*, July 1959, *3*, 211–29; reprinted in E. A. Feigenbaum and J. Feldman, eds., *Computers and Thought*, New York: McGraw-Hill, 1963, 71–108.

Stigler, George J. and Becker, Gary S., De Gustibus Non Est Disputandum," *American Economic Review*, March 1977, *67*, 77–90.

Strotz, Robert H., "Myopia and Inconsistency in Dynamic Utility Maximization," *Review of Economic Studies*, Winter 1955–56, *23*, 165–80.

von Weizsacker, Carl Christian, "Notes on Endogenous Changes of Tastes," *Journal of Economic Theory*, December 1971, *3*, 345–72.

Wilson, Warner, "Reciprocation and Other Techniques for Inducing Cooperation in the Prisoner's Dilemma Game," *Journal of Conflict Resolution*, June 1971, *15*, 167–95.

Winston, Gordon C., "Addiction and Backsliding," *Journal of Economic Behavior and Organization*, December 1980, *15*, 295–324.

Yaari, Menahem E., "Endogenous Changes in Tastes: A Philosophical Discussion," in Hans W. Gottinger and Werner Leinfellner, eds., *Decision Theory and Social Ethics, Issues in Social Choice*, Dordrecht: D. Reidd, 1977, 59–98.

Asset Exchanges and the Transactions Demand for Money, 1919–29

By ALEXANDER J. FIELD*

This paper addresses a general theoretical question—the appropriate specification of the transactions demand for money—as well as a particular historical question: what triggered the Great Depression? Theoretically, fluctuations in the volume and value of asset exchanges in secondary asset markets can influence the transactions demand for money independently of real output and interest rates, and ought to be integrated into the analysis of those forces perturbing[1] the demand for money and shifting *LM* curves in the absence of monetary intervention.

Empirically, I demonstrate that, over the years 1919–29, monthly fluctuations in the volume and value of trading on the New York Stock Exchange did influence the transactions demand for money independently of fluctuations in real output and interest rates. Moreover, in the context of relatively slow post-1925 growth rates of monetary aggregates, the unprecedented increase in the volume and value of such trading from the beginning of 1925 to October 1929 had the effect of shifting the *LM* curve persistently to the left. The failure of U.S. monetary authorities to accommodate this surge in transactions demand, a failure unrecognized at the time, was associated with an antispeculative policy that drove real interest rates to very high levels in 1928–29, levels not approached again until the early 1980's. This deflationary impulse, larger than is apparent from a simple examination of monetary growth figures in relation to *GNP* growth, was the proximate cause of the downturn in real activity generally dated from August 1929.[2]

I. Asset Exchanges and the Transactions Demand for Money

The preponderance of transactions in asset markets involves the conveyance or transfer from one individual to another of previously issued securities. This was true in the 1920's and is true today. A certain base level of transactions results from the issuance of new securities, and from life cycle effects, as individuals liquidate their holdings, or rearrange their portfolios at various stages of their careers. But the observed volatility in trading volume in secondary asset markets is largely driven by other forces, forces related to but distinct from those that determine asset prices. At any moment in time individuals have different expectations about the future course of particular asset prices. Such differences themselves may, but do not necessarily, give rise to trades of existing assets. But when these expectations are changing rapidly in different ways for different individuals as the result of new information differentially available, or differently interpreted, increases in trading volume are likely to result. Such trading volume can persist without necessarily giving rise to changes in an asset price index. By contrast, new infor-

---

*Associate Professor of Economics, University of Santa Clara, Santa Clara, CA 95053. The groundwork for this project was conducted during 1979–80 at the Institute for Advanced Study in Princeton. I am grateful to the Institute for the opportunity afforded to branch into a new area of research. I thank David Laidler and Milton Friedman for detailed comments on earlier versions, and also thank participants in the All-UC Symposium on the Great Depression, Berkeley, May 1982, at which an earlier version of this paper was presented.

[1] Such fluctuations are analogous in their effects to, but analytically distinct from, volatile shifts in liquidity preference: both types of disturbances may cause the demand for money not to be a "stable function" of a "fairly small" number of variables. It is worth recalling that Milton Friedman came to identify the volatility of Keynes' liquidity preference function, not the proposition that the demand for money was interest sensitive, as the distinctive feature of the Keynesian approach (see his 1966 article, p. 85).

[2] This is a statement about the origin—not the depth or duration—of the Great Depression.

mation, widely available and similarly interpreted, may cause asset prices to rise or fall on very low levels of trading.

In other words, prices in secondary asset markets, unlike prices in commodity markets, do not depend for their maintenance on a continual flow of expenditures. Without such a flow, commodity prices eventually decline because of the involuntary accumulation of inventories that ensues in an environment where production flows are positive and the holding of inventories is costly. No such eventuality necessarily accompanies the cessation or slowing of trading in secondary asset markets. Yet, when expectations differ about the appreciation potential of particular assets, exchanges will take place, sometimes accompanied by and sometimes not accompanied by price changes.

If individuals simply bartered financial assets when they wished to trade them, a transactions demand for money could never arise from such exchanges, but for the same reason money serves as a convenience and temporary repository of value in goods transactions, nonmonetary assets tend to be rendered more liquid before being reconverted to other assets. As a result, in periods of high trading volume, a monetary system may experience (at given interest rates and output levels) an increase in the transactions demand for money, especially in large financial centers such as New York City. Trading volume and value, and the resulting transactions demand for money, can therefore be influenced by forces not directly related to flows of production, income, consumption, and saving, and as a result have great potential to move independently of movements in *GNP* or even the average price level of traded assets.

The possible consequences of such movements have been only infrequently examined in the last half century.[3] These and related issues were discussed extensively during the 1920's within the context of a protracted debate over whether or not stock exchange "speculation" deprived other sectors of the economy of stocks of cash balances and/or flows of saving. Many observers, including influential decision makers in the Federal Reserve System, had an intuitive sense that the existence of the stock exchange, and the call loan market in which funds were lent for the purpose of purchasing securities, "drained" funds from the rest of the economy.[4] But advocates of "absorption" employed impressionistic and sometimes contradictory arguments that left them vulnerable to the criticisms of sophisticated monetary theorists such as Gustav Cassel and Fritz Machlup.[5]

The conclusion that the stock market did *not* absorb cash balances was based on a number of irrefutable logical arguments and scattered bits of empirical evidence. Opponents of absorption pointed out correctly that purchases and sales of existing securities in secondary markets did not extinguish cash balances any more than did purchases and sales of a rare painting or a loaf of bread. Critics of absorption went on to emphasize the lack of a close connection between asset prices and the transactions balances necessary to sustain them. Although Machlup waffled on this point,[6] he did claim that there could be in principle no limit to the rise in asset prices over time consistent with a given money supply, so long as new information was interpreted similarly by most investors. He neither emphasized nor denied that the opposite extreme was also possible: high levels of trading among individuals with different and changing expectations about the appreciation potential of different securities in the absence of major movements in individual stock prices and their average.

---

[3] Franco Modigliani, Robert Rasche, and J. P. Cooper (1970) found, but did not report, a very small but positive influence of the value of stock market transactions on the demand for money. This result applied to a regression on quarterly U.S. data from 1955:I to 1966:IV, using a ratio specification. A recent criticism of assuming proportionality between *GNP* and total transactions demand is found in Charles Lieberman (1977), who advocates using total debits, adjusted for currency transactions, as a measure of transactions demand. Two other recent reviewers of research on the demand for money have also noted the potential dangers entailed by the use of *GNP* as a proxy for total transactions demand: "This whole line of reasoning, however, overlooks the large and rapid fluctuations that can take place in the volume of transactions conducted in the financial markets," (David Laidler, 1977, p. 58), and "it is not clear that real *GNP* is a good measure of transactions..." (Steven Goldfeld, 1973, p. 583). Both of these remarks occur in footnotes, however, and are not systematically followed up. Somewhat earlier, Tibor Scitovsky issued a similar warning (1969, p. 39). See also Jacob Enzler, Lewis Johnson and John Paulus, who recognize asset exchanges as possibly influencing transactions demand (1976, p. 269), but empirically are more interested in intermediate goods transactions. Friedman and Anna Schwartz also recognize the possibility that asset exchanges might influence transactions demand (1982, p. 39), although this is not incorporated into their theoretical framework.

[4] See Fritz Machlup (1940, especially chs. 2–4) for a superior although critical introduction to this literature.

[5] For example, the proposition that stock market speculation deprived legitimate commerce and industry of needed capital sometimes coexisted with the proposition that overinflated stock prices made it "too easy" for businesses to raise expansion capital by issuing equities. While these claims can be reconciled by differentiating between those firms that did (because of their size or reputations) and those that did not have access to the stock market, most of the discussion lacked this kind of clarification.

The central issue, recognized as early as 1916 by such German writers as Herbert von Beckerath,[7] and in the 1920's by some of the English and American writers, was whether fluctuations in the volume and value of exchanges of existing securities could have measurable influences on the stock demand for cash balances for transactions purposes. This was never denied as a theoretical possibility. But after extensive debate among monetary theorists in the 1920's, a consensus was reached: such trivial amounts of cash relative to the volume of transactions were involved in the transfer of securities, and the variance in the demand for such balances was so small, due to highly (inexpensively) flexible velocity, that the empirical impact of fluctuations in the volume of asset exchanges on the transactions demand for money had to be insignificant.[8] Even John Maynard Keynes, towards the end of the first volume of the *Treatise*, expressed skepticism as to the empirical importance of a theoretical mechanism he had outlined in some detail at its beginning.[9]

The consensus reached by economists in the late 1920's and early 1930's that fluctuations in asset exchanges had little or no effect on the demand for cash balances has not been systematically reexamined since then, at least insofar as it applies to the 1920's. By the early 1930's, a decade of consistently collected data were available (now assembled in *Banking and Monetary Statistics*, 1976a), but inexpensive statistical techniques for their analysis were not. Applying multiple regression analysis to data from the 1920's, it is now possible to show that over the years 1919-29, the volume and value of trading on the New York Stock Exchange did have an economically important impact on the transactions demand for money in New York and in 100 major commercial and industrial centers, as well as on the demand for more broadly defined national monetary aggregates.

II. Asset Trading and Transactions Demand, 1919-29

According to conventional analysis, the position of an economy's *LM* curve is conditional on a number of parameters, the most important of which are 1) the stock of nominal cash balances in the system, 2) the price level of *GNP*-related transactions, and 3) a given state of liquidity preference. If asset exchanges require fluctuating levels of transactions balances, then at least two additional parameters need to be added: 4) the volume of asset exchanges in secondary markets, and 5) the average volume-weighted price of traded assets. A contractionary open market operation, a change in liquidity preference leading to a higher demand for money as an asset, or a surge in the desired level of asset exchanges can now all, *ceteris paribus*, be interpreted as shifting the *LM* curve to the left, resulting in higher interest rates and lower levels of real *GNP* in the short run. Eventually, price declines in *GNP*-related transactions might lead to counterbalancing shifts in both *IS* and *LM* schedules. In most business cycle episodes, however, various kinds of expectational and institutional inertia keep goods and services prices from moving sufficiently rapidly to prevent quantity movements.[10] Such price movements, and their consequences, will be treated in this paper as of a second order of importance.

---

[6] Compare Machlup "No additional money capital is necessary for a rise in securities prices" (p. 46) with p. 92, where he claims that a lasting stock market boom can result only from inflationary credit supply. Machlup and others dismissed the phenomenon of margin buying as simply an increase in the number of layers of financial intermediation—without real consequence—and pointed to the use of clearance mechanisms and brokerage deposits by regular customers as evidence that velocity was costlessly, or at least very inexpensively, flexible in financial transactions.

[7] "The money which is withdrawn from expenditure on the markets for goods and labour, and used as a unit of account for business on the stock exchange, leads to a temporary reduction in the demand for goods and for labour. That is to say that the money is held up on its way and for the time being can neither be spent nor lent in the economic process proper" (Beckerath, p. 162, cited in Machlup, p. 68).

[8] Machlup (p. 94); Seymour Harris (1933, p. 611); Charles Hardy (1932, p. 167).

[9] Keynes (1934, I. pp. 44 and 249); see my 1983 paper.

Two conclusions follow from these assumptions. A major change in the level of asset exchanges must in the short-run result in some combination of (a) nominal money stock changes, (b) interest rate changes, or (c) changes in real *GNP*. Absent accommodation, the effect will be (assuming the change to be an increase) some combination of higher interest rates and lower real output, depending on the shape of the *IS* curve. With total accommodation, nominal money supply would be higher, with unchanged interest rates and real output. How accommodative can we assume the System to have been?

The Federal Reserve System was set up in part to satisfy temporary seasonal variation in the demand for credit or money associated with such events as fall harvests or December retail sales. The effect of the introduction of the System was to reduce substantially the seasonal variability in interest rates that had previously characterized the American economy.[11] Through a combination of open market operations, foreign exchange transactions, and discount window policy, the Fed's operations permitted the nominal money supply to be endogenously determined by allowing the commercial banking system to satisfy, at least in part, the transactions demand forthcoming, at targeted interest rates, during different months.[12]

The Fed's philosophy with respect to cyclical fluctuations in the demand for credit was less clear. Some argued for accommodating all legitimate "needs of trade," while others maintained that the System should tighten up towards the end of an expansion to prevent speculative "overaccumulation" of inventories.[13] In spite of these and other differences, however, there was fundamental agreement within the System on two propositions: first, that the ultimate source of fluctuations in desired money holdings, be they seasonal or cyclical, were fluctuations in real, *GNP*-related activity, and second, that the responsibility given the System to regulate the supply of money entailed a responsibility to safeguard its purchasing power.

This obligation to maintain the purchasing power of money meant that although the System was willing in the short run to accommodate unexpected cyclical as well as seasonal fluctuations in desired money holdings, originating in perturbations in the *IS* schedule or the money demand function, the willingness to make such accommodation could, over time, come into conflict with the implicit longer-run objective. Periodically, the Open Market Investment Committee, consisting in 1929 of five bank governors with the New York governor as chair, responded to current economic conditions including (at least implicitly) observed longer-run changes in the nominal money supply $(M_t - M_{t-1})$ by ratcheting up or down the interest rate/credit conditions targets. Appropriate changes in discount rate and/or open market purchases or sales consistent with these targets were then made.

---

[10] Examples include overlapping contracts for the delivery of output as well as the purchase of inputs, including labor, or the failure of inflation expectations to adjust downward (upward) until several periods of lower than (higher than) expected inflation have elapsed.

[11] See Friedman and Schwartz (1963, pp. 293–95).

[12] "Transactions demand for money" is used here in the sense of desired money holdings. Such fluctuations could have their origins in the perturbations in the demand for money function (i.e., influences on desired money holdings other than changes in Y or r), or in changes in the flow demand for credit, as reflected in the position of the *IS* schedule.

[13] See Friedman and Schwartz (1963, p. 263). Support for both of these positions is to be found in the System's *Tenth Annual Report*. See Hardy, pp. 79–81.

Ceteris paribus, monetary growth rates would change as a result of such changes. Although primarily concerned with credit conditions and interest rates, policymakers could not help keeping an eye on monetary growth rates, both for what such rates told them about the health of the real side of the economy, and for the warning they might give of longer-run inflationary dangers. Accommodating short-term seasonal or cyclical fluctuations in the desired money holdings was an important objective of the system, but so too was price stability or mild deflation over the longer term, and most central bankers believed the implications of the quantity theory of money: that control of the longer-term growth rate of the money supply in relation to the growth rate of real *GNP* was the necessary means to this end.

In normal times, then, the System can be viewed as having a short-run (accommodative) interest rate target, and a longer-run (less accommodative) monetary growth target, with a feedback mechanism linking the two. The problem was that the late 1920's were not "normal times." In particular, the regressions and simulations discussed below demonstrate that an empirically important increase in transactions demand resulted from the rising volume of asset exchanges in the post-1925 period. Because this increased demand was unrecognized, it short-circuited a feedback mechanism that otherwise would have led to a moderation of the high interest rate policy that pushed the economy into recession in 1929. Before the logic of that argument is developed, however, the influence of asset exchanges on the transactions demand for money must be established econometrically.

The regressions reported in Table 1 test the fundamental empirical question at issue in this paper. Using monthly data, they ask whether, when one holds the level of nominal income and interest rates constant, the level of trading on the New York Stock Exchange exercised a statistically significant and economically important impact on the demand for and holdings of nominal cash balances. Six monetary aggregates serve as the dependent variables in Table I. These include demand deposits in New York City, demand deposits in 100 major centers (excluding New York City), demand deposits in 101 centers (including New York City), currency plus demand deposits in commercial banks (*M*1), currency plus demand and time deposits in commercial banks (*M*2), and *M*2 plus deposits in mutual savings banks and the postal savings system (*M*3). All dependent and independent variables (excluding dummies) have been transformed into natural logarithms to facilitate the calculation of elasticities. Two short-term interest rates are used: that on prime commercial paper, 4–6 month maturity (*PCP*46), and the rate in the call money market for loans on new securities (*CALL*).

Nominal income data is not available on a monthly basis for the 1920's, and three measures have been used to estimate it. One (*YINT*) is a simple linear interpolation from annual data, with endpoints centered in midyear. The second proxy (*YIND*) is constructed using a two-stage procedure. Annual nominal income is regressed on an annual index of industrial production. The resulting regression coefficients are then used to construct predicted monthly income data based on this annual relationship, and monthly seasonally adjusted industrial production data. The third proxy (*YM*) is constructed by multiplying monthly seasonally adjusted industrial production data times a monthly wholesale price index, less farm products. The choice of income proxy does not materially affect any of the conclusions of the paper. To conserve space, only the results using the third proxy are reported, but regressions have been run using each of the three.

The asset exchange variable (*TRAD*1) consists of the monthly volume of trading on the New York Stock Exchange in millions of shares, round lots times an index of common stock prices (1935–39 = 100). The volume series is based on the Exchange ticker, and excludes odd lots, stopped sales, private sales, split openings, crossed transactions and/or "errors of ommission."[14]

---

[14] *TRAD*1 is an imperfect estimate of trading value, inasmuch as the weights used in the stock price index may not have reflected the relative trading volumes of different securities. On the trading volume series, see U.S. Bureau of the Census, Vol. II, p. 999, note on series X-531.

TABLE 1—DEMAND FOR MONEY REGRESSIONS, 1919–29[a]

	Log of Lagged Endogenous Variables	Log of Trading Value, NYSE[b]	Log of Nominal Income Proxy[c]	Log of Short-Term Interest Rate		Seasonal Dummies		Constant
				Call Loan	Commercial Paper	January	August	
Demand Deposits:								
NYC:								
(1) $R^2=.980$, $D\text{-}W=1.87$	.754	.020	.040	−.039		.011	−.020	.180
$\rho=.200$ (2.20)	(17.61)	(5.16)	(2.67)	(−4.81)		(1.70)	(−3.28)	(5.61)
(2) $R^2=.985$, $D\text{-}W=1.85$	.772	.016	.024		−.036	.014	−.018	1.75
$\rho=.264$ (2.99)	(16.14)	(3.94)	(1.47)		(−2.62)	(2.10)	(−2.81)	(4.77)
Outside NYC:								
(3) $R^2=.996$, $D\text{-}W=1.97$	.888	.0077	.033	−.025		.012	−.0034	.832
$\rho=.185$ (2.06)	(50.53)	(4.38)	(4.79)	(−6.42)		(3.76)	(−1.15)	(6.06)
(4) $R^2=.996$, $D\text{-}W=1.97$	.891	.0047	.027		−.039	.013	−.0021	.878
$\rho=.174$ (1.93)	(55.04)	(3.02)	(4.42)		(−7.35)	(4.22)	(−.72)	(6.67)
101 Centers:								
(5) $R^2=.994$, $D\text{-}W=1.83$	.835	.013	.036	−.032		.011	−.010	1.34
$\rho=.164$ (1.73)	(34.76)	(5.83)	(4.44)	(−6.92)		(3.02)	(−2.99)	(6.68)
(6) $R^2=.996$, $D\text{-}W=1.82$	.851	.0087	.026		−.038	.014	−.0093	1.27
$\rho=.267$ (2.95)	(31.11)	(3.73)	(2.84)		(−4.80)	(3.61)	(−2.58)	(5.41)
$M1$:								
(7) $R^2=.9996$, $D\text{-}W=2.19$	.874	.0073	.031	−.015				1.09
$\rho=-.512$ (−6.52)	(50.32)	(5.44)	(5.96)	(−5.80)				(6.94)
(8) $R^2=.9997$, $D\text{-}W=2.19$	.886	.0048	.027		−.023			1.03
$\rho=-.533$ (−7.21)	(56.30)	(3.95)	(5.75)		(−6.67)			(7.08)
$M2$:								
(9) $R^2=.9997$, $D\text{-}W=2.01$	.933	.0069	.020	−.016				.587
$\rho=-.393$ (−4.72)	(118.9)	(6.11)	(5.95)	(−7.71)				(7.84)
(10 $R^2=.998$, $D\text{-}W=2.09$	.941	.0042	.017		−.023			.556
$\rho=-.477$ (−5.92)	(143.2)	(4.48)	(5.76)		(−9.50)			(8.60)
$M3$:								
(11) $R^2=.9998$, $D\text{-}W=2.03$	.944	.0062	.016	−0.14				.505
$\rho=-.404$ (−4.90)	(139.2)	(5.97)	(5.46)	(−7.63)				(7.65)
(12) $R^2=.9999$, $D\text{-}W=2.12$	.951	.0038	.013		−.020			.477
$\rho=-.485$ (−6.08)	(166.8)	(4.38)	(5.18)		(−9.33)			(8.33)

Sources: Federal Reserve System (1976a): Demand Deposits: Tables 49 and 50, pp. 164–74; 196–206; Prime Commercial Paper Rates: Table 120, p. 450; Call Loan Rates, New Loans: Table 120, p. 450; Index of Common Stock Prices: Table 133, pp. 480–81; Volume of Trading on NYSE: Table 135, p. 485. Federal Reserve System (1976b): Seasonally Adjusted Industrial Production: p. S-27. Friedman and Schwartz (1963): $M1$, $M2$, $M3$: Appendix A, pp. 710–12. U.S. Bureau of the Census (1975): Volume and Value of Trading on NYSE, 1935–39, Series X-525—526, Vol. II, p. 1007. U.S. Department of Commerce (1940): Wholesale Prices, All Commodities other than Farm Products, p. 18.

[a] $n=131$ observations (February 1919–December 1929.) The method of estimation was ordinary least squares with first-order autocorrelation coefficient selected by a maximum-likelihood procedure. Durbin-Watson ($D\text{-}W$) statistics are biased towards 2 in such models. However, an approximation of the Durbin h-statistic, $h=(1-d/2)(N/(1-N(\text{var}.\alpha)))^{1/2}$, where d is the $D\text{-}W$ statistic, N is sample size, and α is the estimate of the coefficient on the lagged endogenous variable, enables us to reject at the 5 percent level the hypothesis that the errors in the estimated equation are serially correlated. The t-statistics are shown in parentheses. The t-statistic for ρ is to its right.

[b] Trading volume, millions of shares × asset price index (see text).

[c] Industrial production index, seasonally adjusted, × WPI, other than farm products, × a scaling factor (.0518) to make estimates comparable to annual GNP estimates.

This study employs a widely used, although sometimes criticized,[15] partial-adjustment model estimated through the use of a Koyck transformation. Including the lagged dependent variable on the right-hand side of

---

econometric difficulties of constructing a confidence interval around the long-term elasticity estimates.

[15] See Goldfeld (1973, pp. 581–83). Criticisms center around the proposition that demanders of money never actually hold their desired cash balances, and on the

the equation implies that previous month's cash balances (M_{t-1}) adjusted only partially to current month's desired level (M_t^*). Because regressions are run on demand deposits in New York City as well as in 100 centers outside New York City, this specification facilitates the comparison of the speeds of adjustment of different monetary aggregates.

A maximum-likelihood procedure is used to estimate the autocorrelation coefficient ρ.[16]

In addition to the above mentioned variables, two seasonal dummies were introduced in the demand deposit regressions to deal with systematic variation in the deposit-currency ratio over the year. One might anticipate the need for a positive dummy for December to take account of the large fraction of retail sales that take place during the Christmas season, and the resulting increased demand for money. In the 1920's, however, almost all of the impact of the Christmas season shows up in an increase in the demand for currency, as opposed to demand deposits. It is in January, which typically experiences the largest month-to-month drop in the amount of currency in circulation, that demand deposits rise to unusual levels. As merchants deposit their cash receipts from the Christmas season, and consumers cut back on their currency withdrawals, the deposit-currency ratio rises, and so, temporarily, does the amount of demand deposits. By the end of January, demand deposits begin to fall, as the Fed traditionally contracts the money supply. But average weekly January demand deposits tend to be higher than would otherwise be the case, as the result of this seasonal variation in the deposit-currency ratio. As the regressions show, this increase is statistically independent of variation in the other right-hand variables discussed above. The sign of the coefficient on the August dummy can be explained as the result of a shift in the deposit-currency ratio in the opposite direction.[17]

TABLE 2—LONG-RUN ELASTICITIES OF MONEY WITH RESPECT TO TRADING VALUE[a]

	γ	Long-Run Elasticity
New York City	.246	.081
100 Major Centers	.112	.069
101 Centers	.165	.079
$M1$	.126	.058
$M2$	.067	.103
$M3$	.056	.111

[a] Based on Table 1 estimates. Industrial production index × *WPI* less farm products as nominal income proxy; call loan rate as short-term interest rate.

Note in Table 1 that all the regression coefficients, with the exception of that on nominal income in equation (1), are statistically significant and carry the predicted signs. Regressions including a time trend have been run, but are not reported. Adding a time trend does not greatly change the values or significance levels of any of the reported coefficients, and such a time trend is not itself significant when controlling for the other variables.

Statistical significance is not the same, of course, as economic importance. Table 2 calculates long-run elasticities of money demand with respect to trading value, dividing the estimated coefficient by the implied estimates of γ.[18] Setting aside the combined demand deposits in 101 centers, one can rank the remaining five monetary aggregates in the order of their cumulative sensitivity to trading value fluctuations, as estimated by long-run elasticity: $M3$, $M2$, demand deposits in NYC, demand deposits outside of NYC, and $M1$. However, the speed of adjustment implied by the coefficient on the lagged endogenous variable tends to rise the more narrowly one defines the monetary aggregates, and is the highest for demand deposits in New York City. Although $M2$ and $M3$ eventually experience a larger percentage increase in response to a given percentage increase in trading value, the impact coefficient is approximately three times larger for demand deposits in New York than it is for any of the Friedman-Schwartz monetary aggregates.

---

[16] See Charles Beach and James MacKinnon (1978). This procedure is a variant on the original Cochrane-Orcutt technique.

[17] For all years except the recession year of 1921, the end-of-month currency in circulation figures are higher for August than July (and continue at higher than average levels throughout the rest of each year). The beginning of these cash withdrawals, in the absence of immediate compensatory accommodation, leads to a drop in the deposit-currency ratio which results in particularly low demand deposit holdings in August, independently of variation in nominal income, trading value, or interest rates. The August effect is weaker than the January effect, and is statistically significant for New York and for the combined regressions, but not for the 100 major centers outside New York. See Federal Reserve System (1976a, Table 110, pp. 410–11).

[18] γ is the complement of the estimated coefficient on the lagged dependent variable: $\ln M_t - \ln M_{t-1} = \gamma(\ln M_t^* - \ln M_{t-1})$, where $\ln M_t^*$ represents desired money holdings. The long-run elasticity of nominal income, calculated using a similar procedure, is considerably less than 1 for all of the monetary aggregates. The main explanation for this is that monthly *GNP* is proxied imperfectly, another way of saying that it is measured with error. This tends to bias downward the estimated coefficient on income. Improvements in transactions "technology" due to experience effects over the decade, during which nominal income trended upward, may also contribute.

Over the period in question, annual nominal income fluctuated from a low of $69.9 billion in 1921 to a high of $102.1 billion in 1929, while trading value on the New York Stock Exchange fluctuated from an annualized low of $1.6 billion a year (July 1921) to a high of an annualized rate of $97.8 billion (October 1929).[19] Whereas nominal income's 1921 low was 68.5 percent of its 1929 high, trading value's 1921 low was 1.6 percent of its 1929 high. Almost all of the rise in trading value from its July 1921 trough to its October 1929 peak took place in the post-1925 period. One way to consider the economic importance of the Table 2 elasticities is to ask how much the post-1925 growth in asset trading changed money growth rates, as compared with a hypothetical scenario in which interest rate policy and *GNP* growth took their historical levels, but asset trading remained at its pre-1925 levels. The counterfactual posited, therefore, is one in which trading value remained constant throughout the period at the average of its level during the first 72 months of the sample, but interest rates and nominal income had their actual values.

Dynamic simulations of monetary stocks (i.e., simulations in which the lagged dependent variables are simulated rather than actual values) using the regression coefficients reported in equations (1) and (7) were conducted for demand deposits in New York City, as well as Friedman and Schwartz $M1$. These simulations indicate that, absent the post-1925 surge in asset exchanges, average holdings of New York City demand deposits over the 24-month period 1928–29 would have been 20.3 percent lower than they actually were. Holdings of $M1$ would have been on average 17 percent below their actual levels.[20] These results have important implications.

The decision to raise discount rates as a means of controlling speculative activity on the stock market engendered substantial disagreement within the System.[21] Experience in 1920–21 had led to the advocacy of discount policy to control speculation in commodities, where speculation was understood to mean excessive holding of inventory stocks in anticipation of future price rises. Such holdings made an economy particularly vulnerable to downturns in real activity inasmuch as a change in expectations regarding future price levels or rates of increase could rapidly change the stock of involuntarily held inventories, leading to an immediate downturn in production. The sharp rise in the discount rate in January 1920, and the economy's response to it, gave ammunition both to those opposed to such countercyclical policy, who argued that the rise was too severe, and those in favor, who argued that the rise should have come earlier.[22]

In 1928–29, however, debate concerned the use of discount policy to combat speculation not in commodities (whose prices were dropping), but in financial assets, in particular, equities. Although there are some similarities between commodity speculation and speculation in equities, the potential effect of such speculation on real activity is more attenuated, and therefore the rationale for attempting to use monetary policy to control it is less compelling. In proposing to use high interest rates to moderate an explosion in asset prices in a deflationary environment characterized by rapid real growth, the System was moving into uncharted waters, a fact recognized both by proponents and opponents of such policy. There were concerns among opponents—in retrospect, legitimate—that such higher rates would adversely affect real economic activity, much as discount rate policy in 1920 had helped plunge the economy into one of the shortest but most severe recessions in U.S. history. However, opponents of higher interest rates did not disagree with proponents regarding the desirability of some Federal Reserve action to deal with developments on the stock market: they simply believed that asset prices could somehow be deflated through the use of direct pressure (selective controls on loans for speculative purposes) without affecting real activity. Moreover, both opponents and proponents lacked an awareness of the additional source of transactions demand resulting from asset exchanges.

---

[19] These estimates are obtained by taking the minimum and the maximum value of the *TRAD*1 series (constructed by multiplying a common stock price index by the volume of trading), multiplying them by the average price during the base years (1935–39) of the index, ($26.95) and dividing by 100. Average annual stock prices calculated from U.S. Bureau of the Census, Vol. II, p. 1007, Series X-525 and X-526.

[20] New York City demand deposits' actual 24-month average was $5,379 million. The average for the simulation over the same period was $4,286 million. For $M1$, the corresponding figures are $26,300 million and $21,824 million.

[21] Friedman and Schwartz (1963, p. 254).

[22] Friedman and Schwartz (1963, p. 231).

A plausible case can be made that this ignorance resulted in a tighter monetary policy than otherwise would have obtained. Suppose there had in fact been no such additional source of transactions demand (or, alternately, no rise in the value of exchanges). The simulations described above indicate that had asset exchanges remained at their pre-1925 level, the Fed would have been confronted, already in 1928, with the beginnings of a fairly precipitous decline in monetary aggregates (or, in the case of the broader aggregates, their rate of growth). One cannot predict exactly what the System's response to this would have been, but there is good reason to believe that such developments would have strengthened the hand of those who opposed the high interest rate policy, and that the policy would have been moderated, thus avoiding, or at least making less severe the downturn, centered in construction and automobiles, that was gathering momentum already in the second quarter of 1929.

In the event, there was a complete failure to understand the nature or dimensions of the transactions demand for cash arising from the rising volume of asset exchanges. One can, with the benefit of hindsight, argue that observers should have been puzzled by the fact that the rise in interest rates in 1928–29 had little effect on monetary growth rates. Optimists apparently looked at this evidence and concluded that, in the context of the Wall Street mania, high interest rates were not adversely affecting the real side of the economy, since holdings of cash balances, which they implicitly assumed to be entirely *GNP* related, continued to grow at approximately the same rate as before, contrary to what one would have expected in more normal times.

Unfortunately, real activity was being adversely affected. The steady growth in monetary aggregates did not, according to these regressions, reflect a steady growth in *GNP*-related transactions demand. The interpretation suggested by these regressions is that the rise in interest rates, as theory predicts, led to a general decline in desired money holdings, but that this decrease was more or less compensated for by the surge in transactions demand coming from the stock market, leaving $M1$, for example, virtually unchanged from September 1925 to September 1929.

III. Further Econometric Issues

If the transactions demand for money were almost exclusively associated with *GNP*-related transactions, then it would be reasonable to speak of a *CPI*, *WPI*, or *GNP* deflator as approximating its "price," and deflate both nominal income and the monetary aggregate by such a index. But, if the argument of this paper is correct and transactions in secondary asset markets exercised a nonnegligible impact on the transactions demand for money, this procedure for generating a series of real cash balances is subject to criticism. The income proxy can be deflated by the *WPI* to give us industrial production, and the trading value index can be deflated

TABLE 3—INSTRUMENTAL VARIABLE ESTIMATES[a]

	Lagged Dependent Variable	Log of Trading value NYSE	Log of Nominal Income ($IND \times WPI$)	Log of Call Money Rate	C
M1					
$\rho = -.517\ (-6.619)$	.8924	.0058	.029	−.014	.933
$D\text{-}W = 2.128$	(47.75)	(4.016)	(5.212)	(−5.046)	(5.522)
M2					
$\rho = -.393\ (-4.71)$	.9427	.0055	.0195	−.0144	.5008
$D\text{-}W = 2.087$	(108.9)	(4.296)	(5.468)	(−6.611)	(6.101)
M3					
$\rho = -.384\ (-4.600)$	.9522	.0048	.0163	−.013	.424
$D\text{-}W = 2.011$	(129.8)	(4.170)	(5.369)	(−6.692)	(5.969)

[a] Instruments include lagged dependent variable, constant, and lagged values of logs of trading value, nominal income, and call money rate. Correction for serially correlated errors has been made.

by the asset price index to give us trading volume, but what is the appropriate deflator to use for the stock of money? That obviously depends on how important is the asset exchange component of the transactions demand for money. The more important this component, the larger should be the weight attached to the asset price index in the monthly composite deflator. Were asset prices and goods prices positively correlated over the period in question, the relative weighting of such price indexes would not be empirically of great concern. But, for the 1920's, a series of real balances calculated using a composite deflator is very sensitive to the weighting procedure used, because monthly goods prices (the *WPI*) and asset prices (a stock price index) were *negatively* correlated ($r = -.384$) over the 132-month sample. Small changes in the prior specification of the weighting scheme will produce major changes in the calculated series of real balances. Moreover, that weighting scheme cannot adequately be specified without first reaching conclusions regarding the fundamental questions at issue in this study. These considerations argue, I believe, for a nominal specification, not because people necessarily suffered from money illusion and exhibited demand for nominal balances, but rather because such a specification permits testing of an important hypothesis in a manner that does not require prejudging the outcome of that test.

Regressions were nevertheless run on New York demand deposits, deflating all nominal variables (demand deposits, trading value, and nominal income) by the *WPI* less farm products. These still show economically important and statistically significant estimates of the elasticity of the monetary aggregate with respect to deflated trading value.[23] However, the results are somewhat weaker, and, as indicated, I do not feel that this procedure is theoretically satisfactory, because it assumes that asset prices have no weight in the nominal balance deflator.

The assumption of monthly interest rate pegging is an abstraction—a defensible one—but nevertheless an abstraction. There are legitimate reasons for worrying about whether disturbances in the amount of nominal balances demanded and supplied were totally uncorrelated with interest rates, or indeed with the other right-hand side variables. Table 3 reports instrumental variable estimates, conducted using a procedure suggested by Ray Fair (1970). The instruments include the lagged dependent variable, and lagged values of all other right-hand variables. The estimates in Table 3, designed to correct for possible simultaneous equation bias, do not markedly affect the conclusions of this paper. For example, the elasticity of $M3$ with respect to trading value is .111 using ordinary least squares; the instrumental variable estimates imply an elasticity of .100.[24]

---

[23] For the entire sample, the impact elasticity of deflated demand deposits in New York City with respect to deflated trading value = .023 ($t = 2.94$). If one restricts the sample to the 1921–29 period (108 observations), the elasticity estimate and t-statistic are higher: elasticity = .036 ($t = 4.29$). Right-hand variables in the equations were the lagged endogenous variable, the log of deflated trading value, the log of industrial production, the log of the call money rate, and the two seasonal dummies.

Econometric issues aside, there remain alternate interpretations of these results. There is no necessary connection between the volume of stock market trades and the asset price index, but, as an empirical matter, they were closely correlated in the 1920's ($r = .937$). Given this empirical relationship, one alternate interpretation emphasizes the influence of wealth on the demand for transactions balances. Since wealth grew with the value of securities, these regressions might reflect the fact that people demanded and paid the opportunity cost of carrying higher transactions balances at given (Y, r) combinations, but higher levels of wealth. The statistical results would be due to a spurious correlation of the value of asset *exchanges* with the value of *assets*. A second alternate interpretation argues that the spurious correlation is with the *volume* of exchanges rather than the value of assets. Suppose that a growing volume of asset exchanges reflected increased uncertainty, and that increased uncertainty gave rise to an increased demand for liquidity (shift in liquidity preference). These regressions would then be picking up the fact that the demand for money as an asset, not its demand for transactions purposes, had increased due to increased uncertainty, which also affected the volume of trading.

The best available measure of transactions velocity is the debit-deposit ratio—the ratio of the total monthly flow of debits to deposit accounts to the average stock of deposits held. Both of the above hypotheses imply that high levels of asset trading would have been associated (at any interest rate) with higher holdings of stocks of transactions balances (the denominator of a debit-deposit ratio) without being associated with higher flows of debits (the numerator). Therefore, both of these alternate hypotheses imply that fluctuations in the value of asset exchanges should, *ceteris paribus*, have been *inversely* associated with transactions velocity.

IV. Exchange Trading and Deposit-Turnover Rates

Econometric work on short-term variations in velocity is relatively rare, and that which has been done has primarily considered income velocity as the dependent variable. Such work has concluded that income velocity tends to vary procyclically, and the most widely accepted explanation for this is that interest rates vary procyclically and velocity is positively sensitive to interest rates.[25] Interest rates influence velocity because the higher the rate of return on alternate financial assets, the higher will be the individual and aggregate ratio of expenditures to non- or low-interest-bearing cash balances.[26] This hypothesis should also apply

---

[24] In recent years, a number of papers have been published utilizing econometric procedures to test for "causality," where the direction of causality cannot be specified on a priori grounds and the assumption of instantaneous joint determination, reflected in the use of simultaneous equation methods, is rejected as inapplicable or unrealistic. The results of such causality tests seem to be quite sensitive not only to the choice of lag truncation, but also to the prefilters applied to the data. Such sensitivity seems to be particularly acute in the case of the Sims tests, although the conclusions of Edgar Feige and D. K. Pearce (1979) cast some doubt on the procedures in general. Nevertheless, I did run a multivariate Granger test on the Friedman and Schwartz $M1$ data, logging and first differencing the logs of all variables, and using lags of 18 months. Two regressions were run, the first including 18 lagged values of the dependent variable, income (YM), trading value, and interest rates (call money rate) and a constant, and the second minus the trading value variables. The F-ratio

$$((RSS - URSS)\cdot(n-k))/(URSS\cdot q) = 2.07$$

where $n = 113$, $k = 72$ and $q = 18$. The upper 5 percent point for $F(18,41) = 1.87$. This result is consistent with the Table 1 results.

[25] Over the 1919–29 period, the monthly prime commercial paper rate is negatively correlated with industrial production, although the call money rate does vary procyclically in this sample.

[26] A second hypothesis, advanced by Friedman, calls less attention to interest rates. During the cyclical up-

to regressions using a measure of transactions velocity.

If the results of Table 1 are to be explained using either of the two alternate interpretations suggested above, one would in addition expect to find a negative coefficient on trading variables in a velocity regression. Such a negative coefficient would result from the posited association of asset trading with a growth in the denominator, but not the numerator of the debit-deposit ratio, due in the first instance to the growth in transactions balances at given (Y, r) combinations related to higher wealth levels, for which asset prices were a proxy, and in the second instance to the growth in the asset demand for money associated with uncertainty (for which trading volume was a proxy). In contrast, if the argument of this paper is correct, velocity should covary positively with the volume of trading on the New York Stock Exchange.[27]

The regressions reported in Table 4 have as their dependent variable the debit-deposit ratio. The regressions are run for New York City, for 100 centers outside New York,[28] and for 101 centers, including New York. All three series exhibit strong seasonal regularity, (generally strongest in the outside New York City series), and six monthly dummies are included to capture these effects. A partial-adjustment model is again estimated, although without an autocorrelation correction in equation (2), where it is not needed. The call loan rate is used as the short-term interest variable, except for the outside of New York regressions, where an equation using the prime commercial paper rate (3) is also reported. As before, instrumental variable estimates, using a procedure similar to that described for Table 3, result in only minor variations in coefficient estimates, and in this case are not reported. Calculations of Durbin *h*-statistics, conducted as described in the note to Table 1, again reject at 5 percent the hypothesis of serial correlation in the estimated equations.

The results in Table 4 show unambiguously that trading volume as well as interest rates exercised a positive influence on transactions velocity in the 1920's. They also show that the immediate impact of a surge in trading volume on deposit turnover was three to five times higher in New York City than outside of New York City, depending on which of equations (2) and (3) one uses as comparison. The long-run elasticity of deposit turnover with respect to trading volume

---

swing, he argues, measured income is greater than permanent income (which largely determines the stock demand for money) so that measured income velocity (GNI/M) is high during expansion. Similarly, in the cyclical downturn, measured income is less than permanent income, so GNI/M is low. The stock demand for cash balances is more stable over the business cycle than measured income, leading to procyclical movements in measured velocity. Friedman's hypothesis suggests that income should be positively associated with velocity, independently of interest rates. This hypothesis has been criticized by Laidler on the grounds that transitory increases in income increase wealth, and that "elementary notions" of liquidity preference suggest that money balances might receive a portion of it. See Friedman and Schwartz (1963, pp. 642–43); Laidler (1966, p. 56).

[27] Let η_M = the elasticity of the stock demand for money with respect to trading value, and η_{DEB} the elasticity of the flow of debits with respect to trading values. If $\eta_M = \eta_{DEB}$, then trading value will not affect velocity. Two extreme cases would be (a) $\eta_M = \eta_{DEB} = 0$ and (b) $\eta_M = \eta_{DEB} = 1$. In case (a) neither debits nor deposits held are affected by trading value, and in case (b), equally unrealistic, deposits held rise dollar for dollar with debits. The two alternate interpretations of the Table 1 results discussed at the end of Section II imply that $\eta_M > \eta_{DEB}$, and in particular that η_{DEB} is close to 0. The arguments and evidence of this paper, however, suggest the contrary, that $\eta_{DEB} > \eta_M$, which implies positive elasticity of velocity with respect to trading volume, as Table 4 shows. Nevertheless, the Table 4 results cannot discriminate between $\eta_{DEB} > \eta_M = 0$ and $\eta_{DEB} > \eta_M > 0$. In the former case, that of costlessly flexible velocity in the financial circulation, there would be little reason to expect any impact of trading on the real side of the economy. The arguments of this paper, therefore, depend on both the velocity (Table 4) and the stock demand for money (Table 1) regressions.

[28] Regressions run on data for the 100 centers outside of New York required some transformation of the series presented in *Banking and Monetary Statistics*, because the debit data is reported for 140 centers outside of New York (Table 51), whereas the deposit data is for 100 cities outside of New York (Table 50). However, in Table 55, Federal Reserve Board statisticians have adjusted downward the debit data to make them comparable to the deposit data in order to calculate annual debit-deposit ratios. The implicit annual downward adjustment factors applicable to the debit data were extracted from Table 55 and these annual factors were multiplied by the monthly observations to create a monthly debit series comparable to the monthly deposit series.

TABLE 4—DEPOSIT TURNOVER RATES,[a] 1919–29[b]

	New York City	100 Centers Outside NYC		101 Centers, Including NYC
Right-Hand Variable	(1)	(2)	(3)	(4)
Log of Lagged	.450	.565	.429	.500
Endogeneous Variable	(8.79)	(11.88)	(7.85)	(10.7)
Log of Trading	.170	.032	.051	.118
Volume, NYSE	(9.39)	(4.19)	(5.78)	(8.79)
Log of Industrial	.156	.051	.103	.070
Production	(2.05)	(2.04)	(3.25)	(1.52)
Log of Prime			.150	
Commercial Paper Rate			(6.84)	
Log of Call	.079	.072		.076
Money Rate, New Loans	(3.06)	(6.36)		(4.49)
Monthly Dummies:				
February	−.11	−.14	−.12	−.122
	(−6.94)	(−13.09)	(−12.04)	(−9.35)
March	.12	.088	.071	.105
	(6.13)	(6.98)	(5.39)	(6.77)
August	−.029	−.044	−.045	−.037
	(−2.05)	(−4.20)	(−4.66)	(−3.06)
October	.073	.086	.080	.079
	(4.42)	(8.05)	(7.76)	(5.93)
November	−.05	−.056	−.047	−.051
	(−2.76)	(−5.25)	(−4.38)	(−3.62)
December	.064	.087	.090	.075
	(3.68)	(8.16)	(8.59)	(5.43)
Constant	−.234	.034	−.174	−.079
	(−1.40)	(.606)	(−2.19)	(−.78)
R^2	.892	.92	.898	.924
D-W	1.88	1.92	1.95	.181
ρ	.508		.203	.325
	(6.08)		(2.22)	(3.52)

Sources: (in addition to those for Table 1): Federal Reserve System, *Banking and Monetary Statistics, 1919–1941* (1976): Debits to Deposit Accounts, Table 51, p. 234; Annual Deposit Turnover Rates, Table 55, p. 254.

[a] The monthly sum total of debits to deposit accounts divided by the average weekly (Wednesday) stock of demand deposits. Debit figures for 140 centers outside of New York City have been adjusted downward to make them comparable with deposit figures (see fn. 28).

[b] $n = 131$ observations (February 1919–December 1929). The method of estimation was ordinary least squares, with first-order autoregressive correction in equations (1), (3), and (4). The t-statistics are shown in parentheses.

displays a similar multiple: .31 in New York City, according to equation (1), vs. .072–.074 outside of New York, according to equations (2) and (3). These results cast doubt on the two alternate explanations considered for the Table 1 results.

V. Conclusions

These findings naturally lead one to ask whether such relationships can be generalized to other periods, and one must move cautiously in extrapolating, without adequate empirical work, an effect that I believe can be convincingly demonstrated for the 1920's. Discussions in this section must be viewed as more speculative than the conclusion reached in Sections II and III. Nevertheless, preliminary results for the entire 1919–41 period (275 observations) do show statistically significant although somewhat smaller elasticities of $M3$ with respect to trading value.[29]

---

[29] The impact elasticity of $M3$ with respect to trading value is estimated $= .002$ ($t = 2.98$) over the entire period. Long-run elasticity $= .0525$, based on a γ of .0387. The economic history of the 1930's is so unusual, however, that much more empirical work is necessary before meaningful statements can be made about the impact of asset exchanges on the macroeconomic history of that decade.

Analogous mechanisms may today exercise perturbing influences on the transactions demand for money. For example, maturation of large numbers of All-Savers Certificates in October of 1982 was associated with an unexpectedly large increase in monetary aggregates, presumably the result of an accommodated increase in the transactions demand for cash to be used in exchanging these assets for other assets. A plausible demonstration of such effects for the 1920's opens the door to more sophisticated investigations of such possibilities.

Current claims in the 1980's that financial takeovers, such as that of Conoco by Dupont, absorb cash balances have an air strongly reminiscent of the 1920's, both in the quality of the analysis evident, and in the opposition to such ideas generated among monetary economists. Yet, if the arguments of this paper are accepted, one must conclude that proponents of absorption in that earlier decade were partially correct in their intuition, although for reasons they did not always (especially in the United States) articulate well. The same may turn out to be the case with respect to these current discussions. Similarly, this analysis of the 1920's provides a perspective on the influence of international asset exchanges on the transactions demand for dollars, an influence that was underlined by the rise in monetary aggregates associated with the transfer of assets to Iran under the term of the hostage release agreement.[30] More generally, currency substitutability between dollar, mark, and yen may play a role analogous to that of the New York Stock Exchange in introducing instability in the demand for dollar balances in the 1980's.[31]

Setting aside the research opportunities opened up for subsequent decades, this paper does suggest a number of new interpretations of some key aspects of the macroeconomic history of the 1920's. It explains the Friedman-Schwartz puzzling result that income velocity rises hardly at all towards the end of the 1920's, in spite of record high interest rates.[32] It suggests a way of understanding the relationship between trading on the New York Stock Exchange and the downturn in the German economy in 1928, a relation that has been rejected by those who deny that stock market "speculation" could have "absorbed" funds previously loaned in Germany.[33] Finally, it provides some insight into the proximate causes of the downturn in real economic activity in the United States, usually dated from August 1929.

The U.S. monetary authorities allowed the money stock, by any number of definitions, to grow quite rapidly after the recession of 1920–21. The $M1$ stood at $20.5 billion in January 1922, then rose to $26.2 billion four years later (September 1925), approximately the same level recorded for September 1929. $M2$ and $M3$ also grew at slower rates after 1926. Beginning in 1925, and picking up steam in late 1927, however, trading volume and value on the New York Stock Exchange began to rise to historically unprecedented levels. Confronted by pressure at the discount window to provide more reserves, the

---

[30] *Wall Street Journal*, August 17, 1981, p. 20.

[31] I base this remark on discussion with Ronald McKinnon. See his 1982 article.

[32] See Friedman and Schwartz (1963, Chart 16, p. 197). The resolution is that transactions velocity, which was not proportional to income velocity, rose quite dramatically toward the end of the decade, as one might have expected. This emphasis on the potential influence of fluctuations in the transactions-income ratio may also provide insight into a current macroeconomic issue. Many macroeconomists are puzzled by why velocity has grown so slowly since 1981, more slowly than periods in the 1970's characterized by nominal interest rates in the same range (see *Wall Street Journal*, May 24, 1983). The problem again may be that income velocity is not the right variable. The April 1983 *Federal Reserve Bulletin* (p. A-16) indicated that between 1980 and 1982, debit-deposit ratios in all insured banks rose from 198.7 to 324.2, a 63 percent increase in the space of two years. As of February 1983, that figure had risen even further, to 361.3. This evidence suggests that the demand to hold cash balances for transactions purposes may be higher than policymakers would normally expect, given current levels of real output and interest rates, perhaps explaining policymakers' tolerance of a money growth rate exceeding 12 percent (at an annualized rate) in the first half of 1983.

[33] In a recent discussion of this episode Peter Temin briefly entertained the possibility that the "increased volume of activity on the New York Stock Exchange required an increased quantity of money for transactions purposes..." before rejecting it on the grounds that "it is hard to believe [transactions demand] rose very much" (1976 p. 154).

Fed first tried to clamp on selective credit controls, threatening to refuse discount if loans were for "speculative" purposes.[34]

By and large, the System viewed such loans solely with regard to their impact on asset prices, apparently maintaining that asset prices, like goods prices, required a continual flow of expenditures to sustain them. There was little discussion of the possible role played by such credit expansion in satisfying (if this paper is correct) the rise in demand for transactions balances associated with the unprecedented rise in the volume of financial transactions, an accomodation that was essential if such trading volumes were to continue along with normal *GNP* related transactions. Eventually, the System determined that selective controls were arbitrary and discriminatory, and permitted interest rates to rise to very high levels—the only course of action (outside of major institutional changes) consistent with maintaining a policy of virtually no annual growth in $M1$ in the last three years of the decade.

Some corporations benefited from the ability to raise long-term expansion capital cheaply in the inflated stock market. But many other firms and households began to curtail their investment and consumption demand, especially that for consumer durables such as automobiles,[35] thus helping to precipitate the downturn in real *GNP* signified by the drop in industrial output starting in August. It is reasonable to believe that part of this decline was in response to the very high prevailing interest rates. The persistence of those rates, in turn, can be linked to the System's failure to understand the contribution of the rising volume and value of stock trading to the transactions demand for money.[36] It was in this sense, and really in this sense alone, that one can speak of stock market speculation absorbing cash balances, thereby adversely affecting the real side of the economy.

---

[34] Friedman and Schwartz (1963, pp. 254–66).

[35] See Temin (pp. 64; 171–72) for a discussion of the importance of consumption declines in 1930. However, Temin attributes the decline in consumption to the effect of the stock market crash on household wealth. Yet seasonally adjusted automobile production peaked in June 1929, several months before the crash. I attribute part of the decline in consumer durables production and consumption, as well as the failure of construction to continue to recover after mid-1928, to the tighter credit conditions associated with the Fed's antispeculative policies. See *Survey of Current Business* (January 1930, pp. 2–3, 7, 12, 17, 49, 65). For a more extensive discussion, see my 1984 article.

[36] This argument does not deny the possibility that autonomous forces were generating concurrent contractionary shifts in the *IS* schedule. Nevertheless, high real interest rates prevailing in 1928–29 do suggest that the deflationary impulse came primarily from movements in the *LM* not the *IS* schedule.

REFERENCES

Baumol, William J., "The Transactions Demand for Cash: An Inventory Theoretic Approach," *Quarterly Journal of Economics*, November 1952, *66*, 545–56.

Beach, Charles M. and MacKinnon, James G., "A Maximum Likelihood Procedure for Regression with Autocorrelated Errors," *Econometrica*, January 1978, *46*, 51–88.

Beckerath, Herbert von, *Kapitalmarkt und Geldmarkt*, Jena: G. Fischer, 1916.

Cooley, Thomas F. and Leroy, Stephen F., "Identification and Estimation of Money Demand," *American Economic Review*, December 1981, *71*, 825–44.

Enzler, Jacob, Johnson, Lewis and Paulus, John, "Some Problems of Money Demand," *Brookings Papers on Economic Activity*, 1:1976, 261–80.

Fair, Ray C., "The Estimation of Simultaneous Equation Models with Lagged Endogenous Variables and First Order Serially Correlated Errors," *Econometrica*, May 1970, *38*, 507–16.

Feige, E. L. and Pearce, D. K., "The Casual Causal Relationship Between Money and Income: Some Caveats for Time Series Analysis," *Review of Economics and Statistics*, November 1979, *61*, 521–33.

Field, Alexander J., "Price and Quantity Adjustment in the *Treatise on Money*," working paper, University of Santa Clara, 1983.

______, "A New Interpretation of the Onset of the Great Depression," *Journal of Economic History*, June 1984, *44*, forthcoming.

Fisher, Irving, *The Purchasing Power of Money: Its Determination and Relation to Credit*, *Interest and Prices*, New York: Macmillan, 1911.

Friedman, Milton, *Studies in the Quantity Theory of Money*, Chicago: University of Chicago Press, 1956.

______, "Interest Rates and the Demand for Money," *Journal of Law and Economics*, October 1966, *9*, 71–85.

______ **and Schwartz, Anna J.,** *A Monetary History of the United States, 1867–1960*, Princeton: Princeton University Press, 1963.

______ **and** ______, *Monetary Trends in the United States and the United Kingdom: Their Relation to Income, Prices and Interest Rates, 1867–1975*, Chicago: University of Chicago Press, 1982.

Garvy, George and Blyn, Martin R., *The Velocity of Money*, New York: Federal Reserve Bank of New York, 1969.

Goldfeld, Steven M., "The Demand for Money Revisited," *Brookings Papers on Economic Activity*, 3:1973, 577–638.

______, "The Case of the Missing Money," *Brookings Papers on Economic Activity*, 3:1976, 683–730.

Gordon, Robert J., *Milton Friedman's Monetary Framework: A Debate with his Critics*, Chicago: University of Chicago Press, 1974.

Granger, C. W. J., "Investigating Causal Relations by Econometric Methods and Cross-Spectral Methods," *Econometrica*, July 1969, *37*, 424–38.

Hansen, Alvin, *Monetary Theory and Fiscal Policy*, New York: McGraw-Hill, 1949.

Hardy, Charles O., *Credit Policies of the Federal Reserve System*, Washington: The Brookings Institution, 1932.

Harris, Seymour E., *Twenty Years of Federal Reserve Policy, Including an Extended Discussion of the Monetary Crisis, 1927–1933*, Vol. 2: *The Monetary Crisis*, Cambridge: Harvard University Press, 1933.

Keynes, John Maynard, "Is There Inflation in the United States?," (1928), in Donald Moggridge, ed., *The Collected Writings of John Maynard Keynes*, Vol. XIII: *The General Theory and After, Part I, Preparation*, London: Macmillan, 1973, 52–76.

______, *A Treatise on Money in Two Volumes*, Vol. I: *The Pure Theory Of Money*, Vol. II: *The Applied Theory of Money*, London: MacMillan, 1934.

______, *The General Theory of Employment, Interest and Money*, New York: Harcourt Brace, 1964.

Laidler, David, "Some Evidence on the Demand for Money," *Journal of Political Economy*, February 1966, *74*, 55–68.

______, *The Demand for Money: Theories and Evidence*, 2d ed., New York: Dun-Donneley, 1977.

Lieberman, Charles, "The Transactions Demand for Money and Technological Change," *Review of Economics and Statistics*, August 1977, *59*, 307–17.

______, "A Transactions vs. Asset Demand Approach to the Empirical Definition of Money," *Economic Inquiry*, April 1979, *18*, 237–53.

McKinnon, Ronald, "Currency Substitution and Instability in the World Dollar Standard," *American Economic Review*, June 1982, *72*, 320–32.

Machlup, Fritz, *The Stock Market, Credit and Capital Formation*, trans. by Vera C. Smith, *Börsenkredit, Industriekredit and Kapitalbildung*, 1931, New York: Macmillan, 1940.

Modigliani, Franco, Rasche, Robert and Cooper, J. P., "Central Bank Policy, the Money Supply and the Short Term Rate of Interest," *Journal of Money, Credit and Banking*, May 1970, *2*, 166–218.

Owens, Richard N. and Hardy, Charles O., *Interest Rates and Stock Speculation*, New York: Macmillan, 1925.

Reed, Harold L., *Federal Reserve Policy, 1921–1930*, New York: McGraw-Hill, 1930.

Reisch, Richard, "Rückwirkungen der Börsenspekulation auf den Kreditmarkt," *Zeitschrift fur Nationalökonomie*, 1929, *1*, 205–21.

Rogers, James H., *Stock Speculation and the Money Market*, Columbia: Lucas Bros., 1927.

Scitovsky, Tibor, *Money and the Balance of Payments*, Chicago: Rand McNally, 1969.

Selden, Richard T., "Money: Velocity of Circulation," in *International Encyclopedia of the Social Sciences*, New York: Macmillan, 1968.

Temin, Peter, *Did Monetary Forces Cause the Great Depression?*, New York: Norton, 1976.

Tobin, James, "The Interest Elasticity of Transactions Demand for Cash," *Review of Economics and Statistics*, August 1956, *38*, 241–47.

______, "Liquidity Preference as Behavior Toward Risk," *Review of Economic Studies*, February 1958, *25*, 65–86.

Warburton, Clark, "The Volume of Money and the Price Level Between World Wars," *Journal of Political Economy*, June 1945, *54*, 153–64.

Federal Reserve System, *Federal Reserve Bulletin*, April 1983, Washington: Publications Services.

______, (1976a) *Banking and Monetary Statistics, 1914–1941*, Washington: Publications Services, 1976.

______, (1976b) *Industrial Production, 1976 Revision*, Washington: Publications Services, 1976.

U.S. Bureau of the Census, *Historical Statistics of the United States, Colonial Times to 1970*, Washington: USGPO, 1975.

U.S. Department of Commerce, *Survey of Current Business*, February 1930, *10*; September 1940, *20*.

U.S. House of Representatives, *Federal Reserve Act of 1913, with Amendments and Laws Relating to Banking*, Washington: USGPO, 1966.

Wall Street Journal, August 17, 1981, p. 20; "Velocity Throws the Money Men a Curve," May 24, 1983.

On the Efficiency of Experimental Double Auction Markets

By DANIEL FRIEDMAN*

A double auction is a real world, do-it-yourself market procedure in which participants may make public offers both to buy ("bids") and to sell ("asks"). Experimentalists have for some time noted its "surprising" efficiency; for instance, after listing several aspects of this efficiency in his recent survey article for this *Review*, Vernon Smith remarks:

> Perhaps the most important general feature of the experimental results...is the support they provide for what might be termed the Hayek hypothesis: [Double Auction] *Markets economize on information in the sense that strict privacy together with the public messages of the market are sufficient to produce efficient* *C*[ompetitive] *E*[quilibrium] *outcomes.* [1982, pg. 947].

In his recent survey, Charles Plott comments, "The competitive model seems to work best when markets are organized as oral double auctions.... The overwhelming result is that these markets converge to the competitive equilibrium even with very few traders" (1982, p. 1493). Neither author attempts a theoretical explanation; indeed, at the end of his article, Smith leaves the "Hayekian" or competitive market performance as a "scientific mystery."

The goal of this paper is to solve the mystery, at least in part. Of particular interest is how agents, who may initially be quite ignorant of others' circumstances and strategies, may be guided by the double auction institutions to transact at market-clearing prices.

The issue is important for both theoretical and practical reasons. The Walrasian model that underlies most of our theorizing about competition suggests that a large number (strictly speaking, an uncountably infinite number) of agents, not the handful actually employed in experiments, are required to produce competitive market outcomes (compare Robert Aumann, 1964). Hence these experimental results raise doubts regarding the cogency of the Walrasian model. On the practical side, one notes that many of our most important financial markets, including most U.S. money markets, commodities markets, and interbank foreign exchange and Eurocurrency markets, already are conducted as double auctions. Moreover, the SEC's 1971 Institutional Investor Study recommended that the NYSE abandon its traditional "specialist" institution in favor of an electronic double auction, perhaps similar to that employed by NASDAQ in recent years. If there are sound theoretical reasons to believe that a double auction structure would enhance market efficiency as much in the stock market as it evidently has in laboratory experiments, economists would be able to make strong policy recommendations to the SEC for reform of the NYSE.

It is perhaps surprising in view of their prevalence in the world and in experiments that double auctions have so far received very little attention from theorists. David Easley and John Ledyard (in the only direct theoretical study of which I am aware) point out:

> A participant in a Double Oral Auction has a very complex decision problem. He must decide when to bid [or ask]; how much to bid [or ask] and, whether or not to accept the trades

---

*Visiting Assistant Professor of Economics, Schools of Business, University of California, Berkeley CA 94720. This work was partially supported by the Foundation for Research in Economics and Education, and by the Center for the Study of Futures Markets, Columbia University. I thank Glenn Harrison and Jon Salmon, my collaborators in related experimental work, for helping me focus on the theoretical issues examined here. An early version of this paper was presented at a UCLA workshop, and participants gave useful suggestions. I also owe a considerable debt to two referees for thoughtful criticisms that helped me recast and sharpen the exposition.

> offered by other subjects. Further, all of these decisions must be made with very imperfect information. The subject does not know the payoffs or expectations of other agents, he does not know the terms of trade that will be available to him in the future, and he does not know the effect of his actions on the actions of others. This is a very complex incomplete information game.... [1983, p. 8]

They circumvent many of these complexities by assuming that agents maintain "reservation prices" that adjust over time in plausible ways, and that agents follow bid, ask, and acceptance rules[1] based on their current reservation prices. For a slightly simplified version of a popular experimental market structure, they prove asymptotic convergence of transacted prices and final allocations after many repetitions of the market to (approximate) competitive equilibrium, at least for some parameter configurations. They also test some implications of their model against experimental data.

The results I derive here are generally complementary to those of Easley and Ledyard. I focus on the efficiency properties of a *single* repetition of the market, although I also discuss informally the asymptotic behavior. I take a more general view of the strategies and learning rules that may be employed by participants in a double auction, and rely on an apparently mild consistency condition, phrased in terms of a simple game in normal form, to generate the results. I find that the double auction will generate outcomes that are *ex post* Pareto efficient in some cases and nearly so in others if the consistency condition (called "No-Congestion Equilibrium") is satisfied.

A systematic analysis of market performance requires several steps. One must first characterize the market participants and market institutions. This step is relatively straightforward, and is taken at the beginning of the next section. Next one must derive tractable descriptions of individual actions that arise from the first step. Many approaches are possible here, and I will devote the remainder of the second section to a quick survey of them and to a motivation for my own approach. The third and final step, characterizing the market outcomes arising from the second step, is begun in a very simplified setting (the "basic case") in the third section. In Section IV, the analysis is extended to a setting commonly employed in experimental markets. The final section summarizes the insights arising from the analysis, suggests extensions, and discusses possible applications to experimental data and to contemporary financial markets.

I. Market Institutions and Strategies

The market exists over a continuous finite time interval $[0,T]$. It is populated by a finite number of agents, often referred to as traders, indexed $i=1,\dots,n$. In the cases examined here, there are only two goods, denoted m and x. Good m is the numeraire, referred to as money or cash. The other good x is sometimes referred to as the asset. Each trader i is characterized by

1) *Endowment*, nonnegative holdings (m_i^0, x_i^0) of the two goods, not both zero, at time $t=0$;

2) *Preferences*, represented by a smooth monotone von Neumann-Morgenstern utility indicator U_i defined over *final* $(t=T)$ holdings of the two goods; and

3) *Information*, summarized at each time t by a set consisting of public information available to all traders at time t as well as private information available only to the given trader; these sets will be specified more fully below.

Our market is a *double auction*: at each time $t \in [0,T]$, each trader i sets a nonnegative *ask price* $p_i^a(t)$, representing the amount of cash he is prepared to accept in exchange for a single unit of the asset. At the same time, each trader sets a nonnegative *bid price* $p_i^b(t)$, representing the amount of cash he is prepared to part with to acquire a single unit of x. The "best" ask price $p_M^a(t)=\min_i p_i^a(t)$, and the "best" bid price $p_M^b(t)=\max_i p_i^b(t)$ are referred to as the *market* ask and bid, respectively. Trader i can buy a unit of x at time t by accepting the market ask price; the resulting transaction consists of a unit increase in i's holding of x and a decrease of $p^a_M(t)$ in his cash holding, together with offsetting changes in the holdings of his counterparty, the trader j who held the market bid at t. The case of trader i selling a unit of x to the trader j who holds the market bid is analogous. Transactions that would yield negative holdings of money or the asset for either party are not allowed, and traders' bids and asks are restricted accordingly. Holdings $(x_i(t), m_i(t))$ adjust only through transactions occurring according to these rules.

---

[1] Subject to certain restrictions mentioned below in fn. 5.

It is assumed that new public information at time t consists only of the current market bid and ask prices, and an indication of whether a transaction occurred at either of these prices. Thus each trader at time t has available the (partial) history of these public prices, as well as that for his own prices and transactions. It is also assumed that each trader has private prior information, including at least knowledge of his own endowments and preferences.

It may help to have an example of such an institutional structure. Imagine a dozen widget traders scattered across town, each with a terminal linked to a central computer. Each terminal screen displays current market bid and ask prices (in dollars per widget), as well as the trader's own current bid and ask prices and current holdings. If he needs to refresh his memory, he can also get a display of past prices, his previous trades, and the time remaining in the trading day. Whenever he chooses to do so, a trader can push a button to accept the current market ask. The central computer would then verify that our trader has sufficient cash in his account, and then execute the transaction by updating its records of our trader's holdings of cash and widgets, as well as those of the counterparty, who held the market ask. Similarly, a trader can signal acceptance of a market bid price and the transaction will be consummated as long as he currently has at least one widget. A trader can also revise his current bid or ask prices by typing in the appropriate commands. Traders may not know the number and characteristics of other traders present in the market, nor the identity of their counterparties in transactions. At the end of the trading day, traders can pick up their final holdings of cash and widgets at a central warehouse.

I now turn to the derivation of individual actions. Implicit in the characterizations of traders is the following rationality postulate:

ASSUMPTION A: *Each trader's actions at each time t are chosen to maximize expected utility of final holdings, given information currently available to him, his current holdings, and the double auction rules.*

The actions referred to, of course, are the announcement of one's own current bid and ask prices, as well as the choice of which current market bid and ask prices will be accepted. I note in passing that a relatively simple rule for the latter choice is to pick a maximum acceptable market ask price, $v^b_i(t)$ (and a minimum acceptable market bid price, $v^a_i(t)$) and to buy (respectively, to sell) by accepting the market price whenever it is no greater than $v^b_i(t)$ (respectively, no less than $v^a_i(t)$). If such prices always exist and satisfy the coherency condition $p^a_i(t) \geq v^a_i(t) \geq v^b_i(t) \geq p^b_i(t)$, we say that i follows a *reservation price strategy*.[2]

The term strategy is used advisedly. The effect of one's own action depends crucially on the actions of others. For instance, announcing a bid of $1.00 will have absolutely no effect if someone else currently maintains a bid of $1.10, but in the event that everyone else is currently bidding less than $1.00, it will become the market bid and may result in an immediate transaction. In the latter case, it will provide public information to other participants, which may affect their current expectations and subsequent actions, which in turn may well affect one's own future trading opportunities. Hence once we specify precisely how current information is incorporated into current expectations in Assumption A (a "learning rule"), we have a very complicated continuous-time extensive game of incomplete information.

---

[2] My 1982 paper, Proposition A2, provides a sufficient condition for reservation price strategies to be consistent with the Assumption A. I will not use reservation price strategies in any essential way in this paper, but do refer to them on occasion.

Following the approach pioneered by John Harsanyi (1967; 1968a, b) and more recently articulated by David Kreps and Robert Wilson (1982) among others, one might try to analyze the double auction as follows. First, make decision times and actions discrete,[3] and postulate as common knowledge that traders all believe that others' parameters (preferences and endowments) are drawn from known distributions. Strategies may then be regarded as complete contingency plans, in which the action (or probability distribution over actions) selected by trader *i* at time *t* depends on the public and private information available to him at that time, including in particular the market price histories. One looks for equilibrium selections of strategies and beliefs, for which 1) beliefs (i.e., probability distributions regarding others' parameters, or more germanely, regarding the trading opportunities that others will make available) are updated when new information arrives via Bayes' Rule, using likelihood functions based on others' equilibrium strategies, and 2) strategies are optimal in the sense of Assumption A relative to equilibrium beliefs. Such noncooperative equilibria are variously called Perfect Bayes Equilibria or Sequential Equilibria, according to how one handles events of (equilibrium) probability zero.

This approach to our problem is not very promising for several reasons. The first is its inherent complexity. Drew Fudenberg and Jean Tirole (1983), for example, require a rather intricate analysis of this type even to analyze a market with only two traders, two discrete decision times, and extremely simple initial parameter distributions. One is not encouraged to attempt an extension to more traders, much less to something approaching continuous time. A different sort of simplification, arising from the work of Lloyd Shapley and Martin Shubik, can be found in Pradeep Dubey (1982). One collapses the auction to a game in *normal* form in which strategies consist of one-shot, price-quantity schedules of purchase and sales offers, and outcomes are based on a clearing mechanism. Dubey finds that certain ("active") Nash Equilibria of these games coincide with Competitive Equilibria for the underlying economy under quite general conditions. Leo Simon (1982) reaches a similar conclusion in a related context, emphasizing the role of "Bertrand competition" in forming equilibrium prices. Unfortunately, this work is based on a complete information concept of equilibrium that is inappropriate to our present concerns. Robert Wilson (1982) obtains some results (on *ex ante* efficiency) for a *sealed-bid* double auction using an incomplete information equilibrium concept, but it is not clear how his approach (based on the auction literature following William Vickrey, 1961) can cast any light on how publicly announced bids and offers can provide market participants with sufficient information to achieve outcomes that are efficient *ex post*.

Hence, even heroic simplifications[4] of the explicit incomplete information extended-form game approach to double auctions have so far shown little promise in resolving Smith's mystery. The basic problem is that, within this framework, any learning rule that is fully consistent with Assumption A seems to presuppose an incredible amount of prior information (as well as computational ability!). In the Perfect Bayes Equilibrium concept, for instance, each trader must a priori know, for any realization of parameters, what strategies others would follow. If one asks how such parameter-contingent strategies (themselves information-contingent actions) could possibly be learned if they are *not* known a priori, one is driven to postulate some Nash equilibrium in meta-learning rules, as in Larry Blume et al. (1982). Of course, this only pushes the analysis to a new level of complexity while begging the question of what meta-meta learning process could generate the given meta-equilibrium, and so on, *ad infinitum*.

---

[3] For instance, one might suppose that a computer implementing the double auction has a known cycle time and sequence for polling all traders for acceptances or posted price changes. One can also restrict all prices to integer multiples of some minimal unit, say one cent or a mill.

[4] Another type of simplification, suggested by Glenn Harrison, is to employ Assumption A in a Bayesian learning model in which traders ignore strategic considerations. We are currently investigating the explanatory power of this "Bayesian Games Against Nature" approach. See also Harrison, Smith, and Arlington Williams (1981).

I avoid the dilemma here by an indirect approach. Instead of postulating some explicit "rational" learning rule and a lot of prior information, I seek plausible restrictions on learning and strategies that will yield a tractable and insightful analysis of the double auction for minimal prior information. Of course, I face the danger that if the restrictions are not well chosen, I will obtain only uninteresting tantologies. The reader will have to judge for himself.

Perhaps the best way to begin is to note that there are only four actions that can immediately increase a trader's utility of *current* holdings:

1) accept the market bid ("sell at market") if it exceeds one's marginal valuation (based on current holdings) of the asset;

2) accept the market ask ("buy at market") if it is below one's current marginal valuation;

3) seize ("raise") the market bid if one does not already hold it and it is below one's current marginal valuation; and

4) seize ("shave") the market ask if one does not already hold it, and it is above one's current marginal valuation.

Actions 1) and 2) yield $\Delta U_i(t) > 0$ directly, while 3) and 4) lead to an expected improvement in one's current bundle as long as one believes that there is a positive probability (perhaps quite small) that someone will accept the current market ask or bid. We will say that an agent is *myopic*[5] if he always adopts one of these four actions whenever at least one is available.

However, myopia is not always the best strategy. After all, Assumption A says that traders attempt to increase current expected utility of *final* holdings, rather than utility of current holdings. A trader who believes he knows where prices are going might, for instance, accept a market ask price *above* his current marginal valuation, intending to resell later at a higher price, and thus more than recoup his "short-term loss." It is even possible that a trader may engage in such speculation in order to influence the beliefs of others, as is illustrated in a bit of folklore surrounding Wellington's defeat of Napoleon at Waterloo. The story goes that it was common knowledge in the London stock market on the day of the battle that Nathan Rothschild would be the first to learn the outcome. When Rothschild very publicly sold shares at fairly low prices, other participants inferred that Napoleon had won, and prices dropped precipitously as panic selling ensued. Rothschild's agents then snapped up the shares incognito, providing the wily financier with a large profit.[6]

A third type of behavior, perhaps the most likely if agents are initially quite ignorant of each others' characteristics, is simply to hold back—not accept favorable market prices and not seize the market price—in hopes of transacting at more favorable prices later in the period.

However, as the end of the trading period approaches, the opportunities for gain from holding back or from speculating decline. Traders are then increasingly likely to accept favorable market prices and/or seize the market price, since these are the only ways to actually realize gains in a double auction market. It is possible that some traders might wait too long before employing these myopic strategies and, in a last minute flurry of bids, offers and transactions, some potential gains might go unrealized. However, given any detailed specification of how bids and offers are processed, it seems plausible that traders who know these procedures a priori or have learned them would be able to avoid getting shut off from attempted transactions at the end of the trading round.[7] One does observe in actual experiments that, although there is sometimes a flurry of transactions late in a trading round, experienced subjects seldom miss out on attempted transactions by waiting too long.

---

[5] Easley and Ledyard in effect assume that agents always employ strategies that are "myopic" relative to current *reservation prices* rather than current marginal valuations. The NP-process of my (1979) article formalizes myopic trading in a more general context.

[6] This example also illustrates the point that rational traders need not employ reservation price strategies. Rothschild's acceptance set for market bids must have included a range of prices that were too high to pass up as well as a disjoint range below recent transacted prices that would convey the intended misleading signal.

[7] Indeed, the danger of last-second market congestion clearly will induce rational traders to transact early, particularly those with larger potential gains from a single transaction (see Lorraine Glover, 1983), larger

These considerations lead to my restriction on agents' learning rules and strategies, *no-congestion equilibrium*, a consistency condition that says in a rather weak sense that our agents are able to achieve a Nash equilibrium in their final actions. Roughly speaking, I ask: if the market were unexpectedly held open an extra instant, would anyone definitely wish to change his bid or ask prices, or accept the market bid or ask after all? If not, we have a no-congestion equilibrium.

One can formalize this idea in terms of a hypothetical *extra chance game*, a single instant, simultaneous-move (i.e., normal form) game appended to our double auction game. Strategies in the extra chance game consist of possibly resetting closing bid or ask prices (in particular shaving or raising the market) and possibly accepting the closing market bid or ask. A strategy of neither resetting prices nor accepting at the market is referred to as status quo. Payoffs for the extra chance game consist of changes in the U_i's generated by transactions according to double auction rules from final holdings in the double auction game. We say that beliefs are *nondogmatic* for the extra chance game if each trader believes that acceptance of the market bid by some trader has positive probability (perhaps quite small), and similarly for the market ask.

DEFINITION: *The strategies played in the double auction game constitute a no-congestion equilibrium, if, for some set of nondogmatic beliefs, the status quo strategy maximizes expected payoff for each player in the associated extra chance game.*

Note that this condition appears to be quite weak in that it only requires that there be *some* beliefs that support the closing actions in the double auction game as *ex ante* optimal; the agents' final actions need not constitute a Nash equilibrium *ex post*. However, it is clear that traders would play myopic strategies in an extra chance game, and this turns out to have significant consequences for market efficiency, as will be shown in the next two sections.

II. The Basic Case

Experimental markets typically involve indivisibilities of the asset and "corner solutions" (zero holdings of money or asset), analytical complications that may obscure the nature of our results. For that reason, I first examine in this section a very simple pure exchange economy, referred to as the basic case, that resembles the world studied in intermediate price theory. For the remainder of this section only, the following assumptions will be in force. With regard to agents' preferences, in addition to the maintained assumptions of smoothness and monotonicity, we now also assume convexity (i.e., $U_i^{-1}(c,\infty)\in\mathscr{R}_+^2$ is always strictly convex) and a boundary condition (i.e., $U_i^{-1}(c)$ doesn't intersect the axes). Cobb-Douglas preferences, for instance, satisfy all these assumptions. It is also assumed that initial endowments are all strictly positive, so corner solutions are no longer possible at any time. With regard to auction rules, we allow unrestricted "odd-lot" trading when a bid (or offer) has been accepted: each party to the transaction (say $i=1,2$) picks a $\lambda_i\in(0,1]$ representing the largest fraction of a unit he desires to transact. Then holdings of both parties change by the fraction $\lambda=\min\{\lambda_1,\lambda_2\}$ of the standard double auction transaction specified earlier in the previous section. Thus the asset is regarded as perfectly divisible.

It will be convenient to refer to i's marginal rate of substitution of the asset for money at current holdings, $g_i(t)=\partial_2 U_i(m_i(t), x_i(t))/\partial_1 U_i(m_i(t), x_i(t))$, where ∂_k denotes the kth partial derivative. Note that our maintained assumptions of smoothness and monotonicity ensure that $g_i(t)$ is well-defined and positive. Also note that U_i increases for a small purchase (sale) at price p iff $g_i>p$ (respectively, $g_i<p$), and that by convexity, g_i decreases (increases) for small purchases (sales) at prices near g_i.

---

desired volume, greater risk aversion, or certain types of asymmetric information.

We can now neatly characterize equilibrium states for our basic case. Eschewing the slightly ambiguous term "competitive equilibrium," we have a *price equilibrium* (Gerard Debreu, 1959, p. 93) in a (feasible) allocation (m^*, x^*) and a price p^*, if, for each $i = 1, \dots, n$,

$$g_i(m_i^*, x_i^*) = p^*; \tag{1}$$

that is, the allocation is Pareto optimal and supported by p^*. If we also have for each $i = 1, \dots, n$,

$$m_i^* + p^* x_i^* = m_i^0 + p^* x_i^0, \tag{2}$$

then we have a *Walrasian equilibrium*. Note that the budget condition (2) will generally be violated by trading at the "false prices," (i.e., $p(t) \neq p^*$) that can be expected early in a double auction.

PROPOSITION 1: *If there are $n \geq 3$ traders in a double auction market, and the strategy selection is consistent with Assumption A and results in a no-congestion equilibrium, then the outcome of trading constitutes a price equilibrium: the final holdings are Pareto optimal, and closing market bid and ask prices coincide and support the final holdings.*

PROOF:

A simple arbitrage argument eliminates the possibility that the market bid ever exceeds the market ask. Hence, if the closing market bid p_M^b and closing market ask p_M^a do not coincide, we may assume without loss of generality that these market prices are held by traders 1 and 2, respectively, and $p_1^a > p_2^b$. Letting g_3 denote trader 3's marginal rate of substitution at the final holdings, we have three cases. If $g_3 \geq p_1^a > p_2^b$, then trader 3 could increase his expected utility in an extra chance game either by accepting the market bid (buying from trader 1) until $g_3 \leq p_1^a$, or else (given nondogmatic priors) by seizing the market bit from trader 2. Either way trader 3 will not employ a status quo strategy, so this case is impossible in no-congestion equilibrium. Similarly, $p_1^a > p_2^b \geq g_3$ is inconsistent with no-congestion equilibrium: trader 3 will sell to trader 2 or shave trader 1's ask. But the remaining case $p_1^a > g_3 > p_2^b$ can't occur in no-congestion equilibrium either; trader 3 could gain by buying at a price higher than the market bid or selling at a price lower than the market ask, so he would either "shave" or "raise" the market prices. Therefore, $p_M^a = p_M^b$; call this common value p.

Suppose now that $g_i > p$ for some i. Clearly i does not hold the market bid and could profit by accepting it in an extra chance game until $g_i \leq p_M^b$, so the supposition is false in no-congestion equilibrium. Similarly $g_i < p$ is impossible in no-congestion equilibrium: i should sell at the market ask. Hence $0 < g_i = p$ for all i.

It is perhaps surprising that only three agents are required to yield efficient final allocations under the apparently weak assumption of no-congestion equilibrium. With only two agents, a bilateral monopoly impasse could occur in which buyer and seller both have overly optimistic beliefs and each looks to the other to make price concessions, even in an extra chance game. The presence of third agents breaks the impasse in a double auction market, because then the buyer (or seller) must compete to hold the market bid (or ask) or else be frozen out of the market. Thus, a form of "Bertrand competition" is at work.

A natural question to ask at this juncture is whether the outcome could actually constitute a Walrasian equilibrium. A sufficient condition for Walrasian equilibrium (in addition to the hypotheses of the proposition) is that all transaction prices coincide, but this condition can be ensured by an appropriate choice of reservation price strategies. Specifically, suppose that p^* is a Walrasian equilibrium price, the existence of which is well-known under much weaker conditions than those presently in force (for example, Debreu). If each trader i sets $p_i^a(t) = v_i^a(t) = \max\{p^*, g_i(t)\}$, and $p_i^b(t) = v_i^b(t) = \min\{p^*, g_i(t)\}$, then one can verify[8] that the outcome will be a Walrasian equilibrium and furthermore, that these reservation price strategies constitute a (complete information) Nash equilibrium of the double auction game as well as a no-congestion equilibrium. This example incidentally establishes the consistency of the assumptions of the proposition.

---

[8] See my (1982) paper, Proposition 3.2.

Of course, such strategies are unlikely to be employed unless traders have a lot of consistent prior information. *Stationary repetition* as employed by experimenters since Smith (1962) is a device that provides just such information. The market is run several times, with each agent given exactly the same endowment and preferences in each repetition that he had in previous repetitions. The only difference across repetitions in agents' characteristics is that each has priors that are based on a larger public information set (viz, the price history of previous double auctions) in later repetitions.

Given a specific learning scheme (or suitable restrictions on such schemes), one might argue that under stationary repetition the outcome of a double auction, starting from an arbitrary price equilibrium in the initial repetition, will eventually evolve to a Walrasian equilibrium. For instance, if each trader follows some Bayesian procedure for incorporating public information into his (subjective) distributions of available buying and selling prices, then after a sufficient number of stationary repetitions these distributions will tend to converge, as long as no trader had dogmatic priors. The reason, of course, is that the public information (prices observed in previous repetitions) will eventually outweigh the (possibly diverse) first-repetition prior beliefs in the traders' later-repetition posterior distributions. A second, more delicate step in the argument is that the average subjective distributions (to which the individual distributions tend over repetitions of the market) eventually become concentrated at a point. The idea here is that as beliefs (i.e., probability distributions) converge, the range of transacted prices narrow, and vice versa. A final step, more routine, is to verify that this mutually reinforcing process is consistent only if a Walrasian price is the point of concentration. Unfortunately, this program is difficult to carry out in detail, and I will not attempt it here.

III. Experimental Markets

I now turn from the basic case to a specification of the double auction that is popular among experimentalists, in which the asset is indivisible, payoffs are separable in the asset and money, and traders are specialized as buyers or sellers. Hence in this section I will employ the following assumptions in addition to those of Section II. Preferences are of the form[9] $U_i(m,x)=m+h_i(x)$, where h_i is a nondecreasing function defined for x a nonnegative integer. An agent's endowment either satisfies $m_i^0 \geq h_i(1)$ and $x_i^0=0$, in which case i is called a *buyer*, or satisfies $m_i^0=0, x_i^0 \geq 1$, in which case i is called a *seller*.

For simplicity, I further specialize to the case of single unit transactors, that is, $x_i^0=1$ for sellers, and for each agent there is some $g_i>0$ such that $h_i(x)$ equals g_i if $x>0$, and equals 0 if $x=0$. (In this case, the marginal value of second unit of the asset is zero for everybody, so buyers have no incentive to acquire more than one unit of the asset.)

It is not hard to verify in this context that if a price p^* and allocation (m^*, x^*) satisfy

$$(3)\qquad \begin{aligned} &x_i^*=1 \text{ if } g_i>p^*,\\ &x_i^*=0 \text{ if } g_i<p^*,\quad i=1,\dots,n, \end{aligned}$$

then we have a price equilibrium. If we also have the usual budget condition

$$(4)\quad m_i^*+p^*x_i^*=m_i^0+p^*x_i^0, \qquad i=1,\dots,n,$$

then we have a Walrasian equilibrium. Henceforth, I shall for convenience index our agents so that $g_1 \geq g_2 \geq \cdots \geq g_n$. Then it is clear that Walrasian equilibrium prices p satisfy $g_k \geq p \geq g_{k+1}$, where $k=\Sigma x_i^0 \leq n$ is the total quantity supplied of the asset.

---

[9] In experiments, preferences are typically induced by paying subjects in "real money" for their final holdings of "cash" and asset according to some prearranged schedule that reduces to U_i or a linear transformation of U_i. (For instance, a fixed fee may be deducted from m_i.) Sometimes the rules also prohibit resale or repurchase of assets, a practice I will comment on at the end of this section.

TABLE 1: PARAMETERS FROM BOHM-BAWERK'S HORSE MARKET

Buyers		Sellers	
Agent Index	g_i	Agent Index	g_i
1	£30	4	£26
2	28	5	25
3	26	8	21.5
6	24	11	20
7	22	14	17
9	21	16	15
10	20	17	11
12	18	18	10
13	17		
15	15		

Note: $x_i^0(0) = 0$ for buyers; $x_i^0(0) = 1$ for sellers.

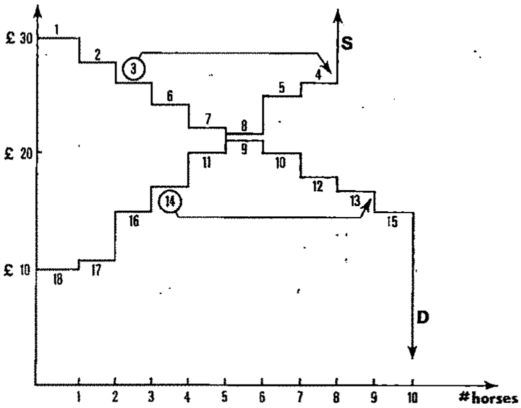


FIGURE 1. SUPPLY AND DEMAND FOR THE HORSE MARKET

A venerable example of such a market from Eugen von Bohm-Bawerk (1891) is reproduced in Table 1, with agents re-indexed according to my scheme. The assets ("horses") are to be traded against cash, denominated in pounds, in a market whose rules Bohm-Bawerk does not fully specify.[10]

Figure 1 displays the staircase-shaped supply and demand curves for this example; the labels attached to the steps indicate the corresponding agent indices. The Walrasian prices consist of the interval [21, 21.5] for which the two curves coincide. One can think of a transaction as reshuffling of the steps—for example, if 3 buys from 14, then as indicated by the arrows in Figure 1, step 3 on the *D* curve moves out to a position of the same height on the *S* curve, and likewise, step 14 of *S* becomes part of the *D* curve, while the gaps in their former positions are closed by sliding the outer portions of the staircases leftward. A transaction can occur to mutual advantage as long as at least one step of *S* remains below its corresponding *D* step. Hence we see that condition (1) above for Pareto efficiency amounts to saying that at the final allocation, the *S* and *D* curves do not intersect in the interior of the first quadrant.

What sort of outcomes might we expect from a double auction in this sort of market? Certainly transacted prices $p(t)$ should be bounded above by g_1 and below by g_n, and we might expect that buyers with low numbers and sellers with high numbers would be more likely to transact early. To the extent that Bertrand competition is effective, the results from the previous section would lead us to expect that the closing prices and allocations would approximate a "price equilibrium." However, after most transactions have occurred, the crucial "third trader" of Proposition 1 may not be so readily available here. For instance, suppose in the numerical example that early transactions gave rise to the configuration illustrated in Figure 2a. Agent 7, the highest valuation buyer remaining, can be expected to hold the market bid, but he can hold it as low as £21 and still outbid 9, his closest rival. Likewise 11 might well hold the market ask as high as $g_8 =$ £21.5. If agents 7 and 11 each thought the probability of a last minute acceptance by the other sufficiently great, an impasse could arise. Impasse seems more likely in the configuration[11] depicted in Figure 2b: agent 8 could hold the market bid as low as £20 and

---

[10] In the first few paragraphs of his argument (p. 204), Bohm-Bawerk seems to be describing an oral double auction. However, in view of his subsequent discussion, a more plausible interpretation is that he envisions an ascending multiple unit auction (in which buyers drop out when the bid price rises above their valuations), together with a descending auction for sellers. He neglects the possibility of "non-truth-telling" strategies. In any case, Bohm-Bawerk predicts a Walrasian outcome.

[11] This configuration could arise if 8 sold (to 1, say) and 9 bought (from 17, say) in early trading.

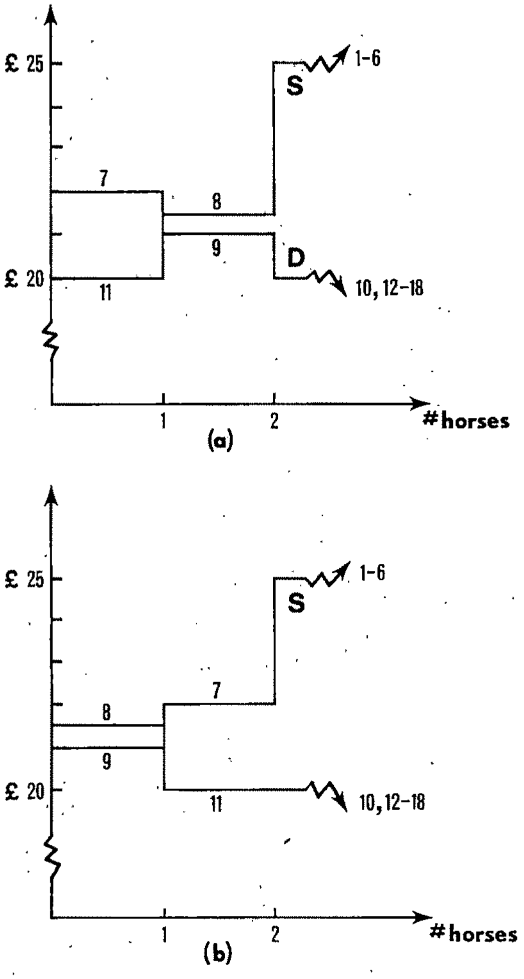


FIGURE 2. INEFFICIENT FINAL CONFIGURATIONS

9 could hold the market ask as high as £22, in which case neither could gain by accepting the market price.

It turns out that such local impasses, involving the four agents that Bohm-Bawerk calls the "marginal pairs," are the worst departure from price equilibrium one can expect in this setting. In making this claim more precise in the next proposition, I will call an allocation (m, x) *almost Pareto optimal* if at most a single transaction is required to make it Pareto optimal. Finally, note that due to ill-conceived speculation in early trading (i.e., buying high and selling low), a buyer could be effectively frozen out of the market by a shortage of cash at the end of trading—say, $x_i(T) = 0$ and $m_i(T) < p < g_i$. This possibility presents no substantial difficulties to this analysis, but it does make my results much more awkward to state.[12] Therefore, it will be assumed for notational convenience that buyers are never cash constrained (i.e., $m_i(t) \geq g_i$ if $x_i(t) = 0$).

PROPOSITION 2: *If the number of sellers* $k \geq 2$ *and the number of buyers* $n - k \geq 2$, *and if traders employ a no-congestion equilibrium strategy selection consistent with Assumption A, then closing market prices satisfy* $g_{k-1} \geq p^a_M(T) \geq p^b_M(T) \geq g_{k+2}$, *and the final allocation* $(m(T), x(T))$ *is almost Pareto optimal.*

PROOF:

Let $J_s = \{i | x_i(T) = 1\}$ and $J_b = \{i | x_i(T) = 0\}$ denote, respectively, potential sellers and buyers at final holdings. Given nondogmatic priors, each seller in an extra chance game would shave the market ask price if it is sufficiently high, so in no-congestion equilibrium, the closing market ask must be held by the seller i_s with lowest g_i, at a price at or below the second lowest $g_i \in J_s$. For similar reasons, the closing market bid price must be at or above the second highest $g_i \in J_b$, and be held by i_b, the buyer with highest $g_i \in J_b$. Hence

$$\begin{aligned}\min{}_2\{g_i : i \in J_s\} &\geq p^a_{i_s}(T)\\ &= p^a_M(T) \geq p^b_M(T)\\ &= p^b_{i_b}(T) \geq \max{}_2\{g_i : i \in J_b\},\end{aligned}$$

that is, the second smallest g_i in J_s is at least as large as the second largest g_i in J_b. Now $J_s = k$ by definition, so $\min_2\{g_i : i \in J_s\}$ is g_{k-1} if $i_b \leq k$, and is g_k otherwise. Similarly, $\max_2\{g_i : i \in J_b\}$ is either g_{k+1} or g_{k+2}, so the claimed bounds on closing market prices follow. If $i_b \geq k+1$ then the closing allocation is actually Pareto optimal by criterion

---

[12] In essence, one has to re-index buyers to reflect "effective" valuations $\min\{m_i(t), g_i\}$, rather than "notional" valuations g_i. See my paper (1982, Section 4) for a treatment of this complication in a related context.

(3) above. Otherwise we must have $i_s \geq k+1$. In this case, if trader i_b were to purchase from trader i_s, then their roles would be reversed (i.e., then $i_b \geq k+1$), so such configurations are almost Pareto optimal; indeed, the single transaction required to achieve Pareto efficiency is that which generates the smallest mutual gains from trade.

The minimal sort of bilateral impasse possible in almost Pareto optimal allocation is not inevitable; by a slight variant of the argument employed after Proposition 1 we can establish the possibility of a Walrasian equilibrium outcome. The only changes necessary are to set a trader's ask price to a prohibitively high level (say $B = \Sigma_i m_i^0$) when he has no units of the asset, and to require that a trader's bid price never exceed his current cash balance.

The learning requirements for achieving something close to Walrasian equilibrium in stationary repetition of an experimental double auction actually seem less severe than in the basic case. For example, it seems plausible in a given repetition that a trader would refuse to buy at prices above the closing ask, or sell at a price below the closing bid, of the previous repetition. In the present context, such policies are quite consistent with no-congestion equilibrium strategies, and enable traders to improve upon their least favorable transactions without reducing trading volume. If all traders follow such policies, there will be a definite tendency for transacted prices to lie in narrower intervals in later repetitions. One presumes that the possible bargaining problem between i_b and i_s would eventually be resolved to mutual satisfaction, in which case final allocations in subsequent repetitions would be Walrasian, and transacted prices would lie in the Walrasian interval $[g_{k+1}, g_k]$. Again, I will not attempt to make this argument rigorous here.

To summarize, for the simple case of a popular experimental design considered in this section, I have established theoretically that closing prices and allocations will be almost efficient even in the first repetition of an experimental double auction market, and have suggested that they will ultimately converge to Walrasian equilibrium values under stationary repetition. These conclusions did not require that traders initially possess any information on others' circumstances or that they employ sophisticated learning rules. Neither did they require that agents be numerous: two traders on each side of the market suffice. Thus I have provided a partial theoretical explanation of the empirical regularities noted in Smith's (1982) Propositions 4, 5, and 6.

I close this section with a few caveats. If agents may transact more than one unit of the asset, the theoretical results are not quite so strong. For some parameter values, a "marginal" agent may have the incentive to act as a local monopolist, and outcomes may not then be quite Walrasian.[13] However, for parameter values commonly chosen in actual experiments, there are no such incentives, and my conclusions remain valid. If repurchases and resales of the asset are prohibited (see fn. 9), then the bounds on closing prices in Proposition 2 may be somewhat looser. In terms of a *S-D* diagram such as Figure 1, transactions may be represented in this case by *removing* the transactors' steps from the diagram, rather than reinserting them to the right. As a result, the "final configurations" as in Figure 2 may have second steps of the *S* and *D* curves (representing the closing price bounds) that are further apart. The indexing is more complicated in this case, but the reasoning is unchanged.

A final caveat is that experimental subjects are human, and may blunder or find ways of colluding. For reasons that should be clear from this analysis, these phenomena should quickly become less important in a double auction market as the number of subjects increases beyond the minimal two on each side of the market. Likewise, the assumption of no-congestion equilibrium seems increasingly plausible, as traders gain experience in a double auction market.

IV. Discussion

The results of the previous two sections illustrate the power of three institutional features of experimental markets in shaping efficient outcomes. The primary feature is the double auction structure: in order to realize any gains from exchange, a trader must either seize the market price, or accept the market price of another. This necessity both limits each agent's influence on prices (via Bertrand competition) and also conveys high quality information to other agents. By contrast, under the "tatonnement" institution implicit in most discussions of price formation (in which each agent provides an excess demand schedule to an outside "auctioneer," who then computes and announces a market-clearing price), agents may be able to influence prices significantly by conveying misleading information (i.e., false excess demands) unless each agent is quite "small" relative to the size of the market.

---

[13] See my paper (1982, Appendix) for an example of this phenomenon in a related context.

The second feature is the preset time T at which trade will end. As the deadline T approaches, agents must become more myopic if they wish to realize any remaining gains from trade, thus enhancing the informational and competitive aspects of the double auction. Indeed, Proposition 1 suggests that these two features alone may suffice to bring about an efficient final allocation (or almost efficient, for Proposition 2) even if agents have very little prior information about each other.

The last feature is stationary replication, a device that makes agents' prior information both more consistent and more precise. I argued informally that under stationary replication, agents need not be very intelligent to learn enough to achieve a Walrasian outcome, for which all trades occur at a single price that clears the market. By contrast, such learning may be impossible in tatonnement markets, as argued in a different context by Roman Frydman (1982), for instance.

There seem to be no obstacles to extending arguments of the sort employed in Sections III and IV to more general cases. For instance, my 1982 paper examines a constant marginal rate of substitution specification employed[14] in experimental asset markets. The case of uncertainty (state-contingent utility functions) is analyzed in an Appendix to my paper with Harrison and Salmon (1983). The extension of the basic case of Section III to a many-asset market seems quite straightforward, at least if preferences are convex. One could easily justify an appropriate convexity assumption under (for instance) a mean-variance approach to asset valuation in which agents do not change their beliefs about the joint distribution of long-run asset returns during the trading process. Direct analogues of my present results appear valid in this case, although (due to lack of a boundary condition) one must qualify the three-trader assumption of Proposition 1 by three *active* traders, that is, for each asset, there are three traders who have positive holdings of cash and the given asset. Of course, when odd-lot trading is not permitted, one can't generally obtain equalized marginal rates of substitution, and the results are more awkward to state in this case. Nevertheless, it seems reasonable to conjecture that, once one has defined Pareto optimal and Walrasian equilibrium with appropriate indivisibility constraints, my results would still go through.

There are other ways to follow up the work presented here. First, it suggests new ways to analyze experimental data, in particular, the unaccepted bids and asks that have usually been ignored (Easley and Ledyard being a notable exception). One could postulate simple specific learning rules—for example, no learning (myopia), or the Bayesian Game Against Nature rules of Harrison, Smith, and Williams. The sequences of observable market bids, and asks and trades arising from Assumption A and reasonable priors could then be generated and compared to the experimental data.

Finally, we may glean new insights regarding price formation in contemporary financial markets. Although the trading procedures of such markets have increasingly come to resemble those postulated in my double auction model, one hardly expects stationary repetition in ongoing markets; to the contrary, new outside information (both public and private) typically arrives *during* the trading process. The end of a trading day, T, is also usually much less important in this context, because participants are generally concerned with payoffs over a longer time span. Hence our results (extended perhaps to the multigood case) do *not* demonstrate that financial markets are efficient; indeed, to the extent that our conditions are violated, they may suggest reasons for the surprising degree of volatility in such markets.

---

[14] See Robert Forsythe, Thomas Palfrey, and Plott (1982), for instance.

REFERENCES

Aumann, Robert, "Markets with a Continuum of Traders," *Econometrica*, January 1964, *32*, 39–50.

Blume, L. E., Bray, M. M. and Easley, D., "Introduction to the Stability of Rational Expectations Equilibria," *Journal of Economic Theory*, April 1982, *26*, 313–17.

Bohm-Bawerk, Eugen von, *Positive Theory of Capital*, 1891; translated by W. Smart, Steckert: New York, 1923 edition.

Debreu, Gerard, *Theory of Value*, Cowles Monograph No. 17, New Haven: Yale University Press, 1959.

Dubey, Pradeep, "Price-Quantity Strategic Market Games," *Econometrica*, January 1982, *50*, 111–26.

Easley, David, and Ledyard, John, "A Theory of Price Formation and Exchange in Oral Auctions," Department of Economics Working Paper No. 249, Cornell University, rev. January 1983.

Forsythe, Robert, Palfrey, Thomas and Plott, Charles, "Asset Valuation in an Experimental Market," *Econometrica*, May 1982, *50*, 537–68.

Friedman, Daniel, "Money-Mediated Disequilibrium Processes in a Pure Exchange Economy," *Journal of Mathematical Economics*, March 1979, *6*, 149–67.

______, "Price Formation in Double Auction Markets," Working Paper No. 278, University of California-Los Angeles, November 1982.

______, **Harrison, Glenn and Salmon, Jon,** "The Informational Efficiency of Spot and Futures Markets," CSFM Working Paper, Columbia Business School, June 1983.

Frydman, Roman, "Towards an Understanding of Market Processes," *American Economic Review*, September 1982, *72*, 652–68.

Fudenberg, Drew and Tirole, Jean, "Sequential Bargaining with Incomplete Information," *Review of Economic Studies*, April 1983, *50*, 221–47.

Glover, Lorraine, "Microdynamic Price Determination," doctoral thesis proposal, University of California-Los Angeles, 1983.

Harris, M. and Townsend, R. M., "Resource Allocation Under Asymmetric Information," *Econometrica*, January 1981, *49*, 33–64.

Harrison, Glenn, Smith, Vernon and Williams, Arlington, "Learning Behavior in Experimental Auction Markets," University of California-Los Angeles, November 1981.

Harsanyi, John, "Games with Incomplete Information Played by Bayesian Players," Part I, *Management Science*, November 1967, *14*, 159–82.

______, (1968a) "...," Part II, *Management Science*, January 1968, *14*, 320–34.

______, (1968b) "...," Part III, *Management Science*, March 1968, *14*, 486–502.

Kreps, David and Wilson, Robert, "Sequential Equilibria," *Econometrica*, July 1982, *50*, 863–94.

Plott, Charles R., "Industrial Organization Theory and Experimental Economics," *Journal of Economic Literature*, December 1982, *20*, 1485–527.

Simon, Leo, "Bertrand and Walras Equilibrium," Working Paper, University of California-Berkeley, 1982.

Smith, Vernon L., "An Experimental Study of Competitive Market Behavior," *Journal of Political Economy*, April 1962, *70*, 111–37.

______, "Microeconomic Systems as an Experimental Science," *American Economic Review*, December 1982, *72*, 923–55.

Vickrey, William, "Counterspeculation, Auctions, and Competitive Sealed Tenders," *Journal of Finance*, March 1961, *16*, 8–37.

Wilson, Robert, "Double Auctions," unpublished manuscript, Stanford University, 1982.

Gambles and the Shadow Price of Death

By JOHN M. MARSHALL*

Recent papers on the "value of life" involve preferences of the sort introduced by Milton Friedman and Leonard Savage (1948).[1] Given the choice between undergoing a particular gamble in wealth and possessing the mathematical expectation of the payoffs of the gamble, the consumers in these studies may prefer the gamble. The presence of such risk-loving preferences is inferred from two well-established facts: life is an indivisible commodity in the models, and indivisibility leads to preference for gambles. The former fact is made clear by Philip Cook and Daniel Graham. The latter was first noted by Yew-Kwang Ng (1965) and is widely known. In spite of the obvious connection, the implications of risk-loving tastes in this area have not been explored.

This paper examines some implications of such risk loving. The idea that indivisibility produces Friedman-Savage preferences and that such preferences create a demand for gambles is extended in Section I to a theory of demand for indivisible goods and derived demand for associated optimum gambles. This theory suggests that the price of an indivisible good will not compensate for possessing or not possessing the good. Section II applies the demand theory to an equilibrium model of labor supply that is relevant to the determination of the value of life. In this model, the wage differential is useful in guiding policy toward public risks, in spite of the fact that the differential is not compensating.

I. Demand for Gambles and Demand for Goods

Consider a continuously divisible good x and an indivisible good z. The consumer's choice set is defined only for $z=0$ and $z=1$. Good z may be one's life, for example, or any other indivisible commodity.[2]

Let $U(z, x)$ represent von Neumann-Morgenstern utility. Assume risk aversion in good x for fixed values of z; that is, $U_2>0$ and $U_{22}<0$ for all z and x, where subscripts denote first and second derivatives.

Good x is the numeraire and the price of z is p. Then the indirect utility of wealth I is defined by

$$V(I,p)=\operatorname{Max} U(z,x), \tag{1}$$

subject to

$$pz+x=I. \tag{2}$$

This may be rewritten as

$$V(I,p)=\operatorname{Max}[U(0,I),U(1,I-p)]. \tag{3}$$

Many different shapes of V are possible. Only cases that lack discontinuities or multiple switches are considered here.[3] An example is given in Figure 1.

The indirect utility function in Figure 1 shows the utility available from wealth held before the person enters the market for goods. Wealth I is a critical point. To the right of I, optimum utility is attained by purchasing the indivisible good at price p and spending the remaining wealth on the divisible good. To the left of I, the indivisible good is not purchased. At wealth equal to I the person is

---

*University of California, Santa Barbara, CA 93106. I thank Brian Ellickson, Tom Cooley, Robert Deacon, Mark Feldman, H. E. Frech, Steve LeRoy, John Riley, and Sherwin Rosen for useful comments.

[1] An exhaustive listing of sources on the value of life is impossible here. M. W. Jones-Lee (1976) reviews many early contributions to the theory. A recent contribution to the conceptual understanding of the value of life is Philip Cook and Daniel Graham (1977). The paper by Richard Thaler and Sherwin Rosen (1976) inspired much of the recent empirical research. Martin Bailey (1980) summarizes some of the empirical work and gives a substantial bibliography.

[2] See the Appendix for examples of indivisibility that have appeared in recent literature.

[3] Other examples exist that follow Friedman-Savage in lacking concavity. Multiple nonconcavities of the indirect utility function are possible, as are discontinuities. Trivial cases exist that preserve the usual concavity: p may be zero, the utility function may not depend upon z, or z may be a bad instead of a good.

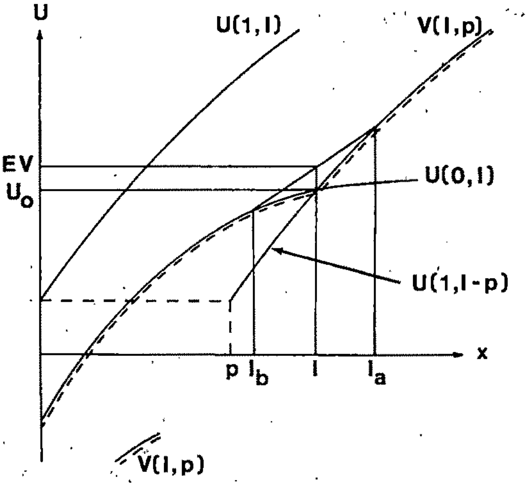


FIGURE 1. CONSTRUCTION OF THE INDIRECT UTILITY FUNCTION AND A BENEFICIAL GAMBLE

indifferent between consuming or not consuming the indivisible good. In either case, utility is U_0.

This indirect utility function implies that some gambles in wealth are preferred to their expectations. For example, consider wealths I, I_a, and I_b in Figure 1. Rather than having wealth I with certainty, the consumer prefers a gamble that yields I_a with probability q and I_b with probability $1-q$, where q satisfies

$$qI_a+(1-q)I_b=I. \tag{4}$$

This gamble improves *ex ante* expected utility from U_0 to EV in Figure 1. For a person possessing wealth I, this gamble is fair in the sense that the mathematical expectation of gain $q(I_a-I)+(1-q)(I_b-I)$ is equal to zero.

The prizes I_a and I_b in the gamble for wealth are associated with particular consumption bundles. The person purchases the indivisible good if he wins the gamble (I_a), but he prefers not to buy it if he loses (I_b). The remainder of wealth is spent on the divisible good. Therefore, the prior gamble in wealth provides the same utility gain that would be supplied by a particular gamble in bundles of commodities.

The gamble of I_a and I_b in Figure 1 is an example of the class of gambles in wealth

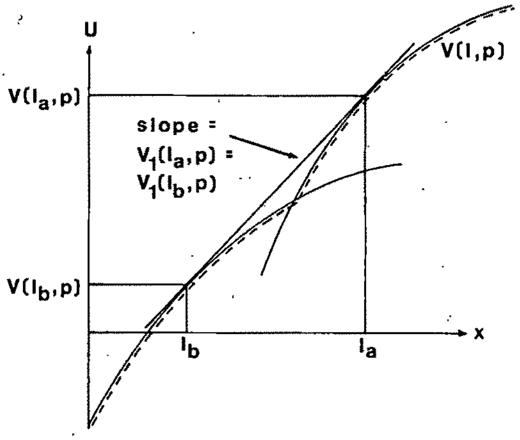


FIGURE 2. THE OPTIMUM GAMBLE

that are preferred to available nongambling alternatives, but it may not be the best of that class. The optimum gamble can be defined for any initial wealth I. The optimum gamble maximizes

$$qV(I_a;p)+(1-q)V(I_b;p) \tag{5}$$

over I_a, I_b, and q subject to the fairness requirement of equation (4). Graphically, the optimum gamble is found by placing a straight line so that it lies tangent to both lobes of the indirect utility function, as in Figure 2. The I_a and I_b are determined by the points of tangency. Thus,[4]

$$V_1(I_b;p)=V_1(I_a;p), \tag{6}$$

$$V(I_a;p)=V(I_b;p)+V_1(I_b;p)(I_a-I_b). \tag{7}$$

These equations may be rewritten in terms of the direct utility functions as

$$U_2(0,I_b)=U_2(1,I_a-p), \tag{6'}$$

$$U(1,I_a-p)=U(0,I_b)+U_2(0,I_b)(I_a-I_b). \tag{7'}$$

---

[4] The model could be expressed in purely algebraic terms. In that case, these expressions would be derived from the first-order conditions. Examples of constrained maxima of this type are given in Section II.

Equations 6 and 7 define optimum I_a and I_b independently of I. Then q is given by equation (4), provided that I lies between I_a and I_b. When I lies below I_b, optimum q is zero: the person chooses not to gamble and not to buy the indivisible good. When I is above I_a, optimum q is unity, and the consumer buys z without making a prior gamble. Gambles involving more than two prizes may also be considered. It is evident that if a gamble with three or more prizes is optimum, there is a two-prize gamble with equal expected utility at the given endowed wealth I. Thus it is sufficient to search for the optimum over gambles with two prizes.

The optimum gamble depends upon the price of the indivisible good. Given an increase in p, the upper branch of the indirect utility function—that part determined by $U(1, I-p)$—shifts to the right, the cusp in the utility curve moves to higher incomes, the tangent line between the two lobes flattens, and the tangency points that define the prizes in the optimum gamble move to the right. For a fixed wealth, increased p changes the odds in the optimum gamble. At low prices, it is preferable simply to purchase the indivisible good. At some higher price, gambles begin to be preferable. At first, the preferred gamble has a high chance of gain, but as prices increases, the chance of gain falls. Above a certain critical price, the person neither gambles nor consumes the good.

The sum at stake in the optimum gamble, $I_a - I_b$, may be large. Suppose that $U_2(0,x) < U_2(1,x)$ for all x. This means that the marginal utility of the divisible good at a given consumption of that good is greater if the person also consumes the indivisible good. The consumption of the indivisible good enhances consumption of the divisible good.[5] When the indivisible good is life, this assumption is certainly acceptable. In other contexts, it may or may not be reasonable. It is used here to illustrate the possible magnitudes of the gambles involved. Since $U_{22} < 0$ and $U_2(0, I_b) = U_2(1, I_a - p)$, we have $I_a - p > I_b$ or $I_a - I_b > p$. The difference between winning and losing exceeds the price of the indivisible good. As a consequence, the winners at the gamble in wealth enjoy higher consumption of both goods than do the losers. This is natural since having the indivisible good raises the marginal value of the divisible one.

II. Value of Life

This section is concerned with expenditure on public safety and with allocation of workers among occupations having differing risk. The attractive idea is that the difference in wages in more- and less-risky occupations can guide the public decision concerning safety. This notion seems to derive from the conclusion of Adam Smith that wage differentials compensate utility losses from less agreeable jobs.[6] The analysis of Section I suggests that prices will not generally compensate in the way required. That suggestion is confirmed below, showing that Smith's conclusion is inapplicable in the case of risky labor. This finding would seem to invalidate the proposed use of wage differences. However, closer investigation will show that compensating differentials are not needed in the argument. What is needed is a sufficiently complete regime of markets. The needed markets are markets in gambles. It will be shown that in the absence of such markets, the use of wage differences to value risks to life is not justified.

The model described below has two occupations of differing risk. Wages are endogenous and workers are free to choose between industries. Consumer-workers may, if they choose, undertake gambles in wealth prior to selecting jobs. Various aspects of the model are taken up in the following order: technology of firms; markets in gambles and insurance; preferences of consumer-workers; character, existence, and interpretation of equilibrium; and optimum expenditures to reduce public risk.

---

[5] The divisible and indivisible goods are substitutes in the usual sense that the cross-partial derivative of compensated demand is nonpositive. The notion that one good enhances consumption of the other does not imply that the goods are complements.

[6] "The whole of the advantages and disadvantages of the different employments of labour and stock must, in the same neighbourhood, be either perfectly equal or continually tending to equality" (*The Wealth of Nations* p. 99).

A. *Firms*

Consider two industries in which deaths of workers, as well as their labor, are inputs to production. In each industry, the proportion between deaths and labor input is fixed. Industry S imposes risk of death s (safe), where s is the proportion between deaths and labor force, and industry R imposes risk of death r (risky), where $r > s$. There is a probability t of nonindustrial death so the whole risk to workers in the two industries is $s+t$ or $r+t$, depending upon occupation.[7] The goods produced by the two industries are perfect substitutes, and the essentially singular good produced by this economy is taken as numeraire.

Define the unit for measuring labor in such a way that the total labor force is one unit. The indivisibility of workers is not an issue here, so workers are regarded as a continuum. Labor inputs are l_s and l_r into industries S and R. Industry S has production function $f(l_s)$ and industry R has production function $g(l_r)$. Assume $f', g' > 0$, $f'', g'' \leq 0$, $g' \to \infty$ as $l_r \to 0$, and $f' \to \infty$ as $l_s \to 0$.

Wage rates in the industries are w_s and w_r. Markets for goods and labor are competitive. As a consequence, profit maximization by all firms implies

$$f'(l_s) = w_s, \tag{8}$$

$$g'(l_r) = w_r. \tag{9}$$

B. *Markets in Gambles and Insurance*

1. *The Market in Gambles.* Since there is an indivisible good, namely life, consumers will demand gambles in wealth. Such demands are satisfied in this model in a market for gambles. The characteristics of this market depend upon the fact that gambles are contracts. The parties to a gamble agree upon certain outcomes, upon the probabilities of the outcomes, and upon the transfers to be made between the parties in each outcome. Although the terms of the contracts are governed by market forces, resources do not change hands in the market for gambles. The outcomes of the gambles are generated later by a random device that conforms to the agreed probabilities. At that time, the parties to the gamble transfer ownership of resources according to the terms of the contract.

Neither moral hazard nor adverse selection is present in the market for gambles. This is a consequence of setting the probabilities as part of the contract. In particular, adverse selection is absent because probabilities are not concealable. This feature may be best appreciated by comparing the present market to that for life insurance. In life insurance, adverse selection may exist because the probability of death is a characteristic of the demander that the demander can sometimes conceal from the supplier. Such concealment cannot occur in the market for gambles. Similarly, moral hazard is absent because the demander is unable to affect the probabilities after the contract is made. The reasoning applied above to demanders applies equally well to suppliers of gambles. Suppliers cannot hide or change the agreed probabilities, and therefore suppliers' actions cannot lead to adverse selection or moral hazard.

A gamble is characterized by the prizes I_a and I_b, and the corresponding probabilities q_a and q_b, where $q_a + q_b = 1$. Gambles with more prizes are not required in this analysis although they could be allowed.

2. *Supply of Gambles.* The supply of gambles is assumed to be competitive, risk neutral, and free of transactions costs. As a consequence, suppliers earn neither a transactions fee nor a risk premium. Since they are risk neutral, suppliers are indifferent concerning the probabilities involved in the fair gambles they supply. Thus each demander is free to choose from among all possible fair gambles in wealth the one that is personally optimal.

---

[7] The events "industrial death" and "nonindustrial death" are assumed to have null intersection. This reduces the complexity of several equations appearing here without real loss of generality.

These assumptions about supply are partly justified by the following considerations: since there is a continuum of workers, there are many demanders of any kind of gamble. Thus it is possible for supply to be accomplished by a literal lottery. That is, a fixed number of winners is chosen at random from the pool of entrants. In a lottery, the aggregate value of prizes is perfectly certain and equal to aggregate receipts. Under this style of supply, there is no risk in the supply of gambles and no need to compensate the suppliers by paying a risk premium.

3. *Supply of Insurance*. Workers will demand insurance for the risk of death. Fair insurance is supplied in this model. Since workers are a continuum, the aggregate expenditures for the supplier of insurance may be forecast with certainty, thus there is no risk premium in the price of insurance. Assuming competition, absence of transactions costs, and no taxes, the supply of insurance is fair. These considerations are parallel to the considerations that lead to a costless supply of fair gambles. A more elaborate justification of fair prices could be based upon the work of Edmond Malinvaud (1972).

The equilibrium described below includes insurance, but the assumed nature of supply of insurance makes it unnecessary to treat the insurance sector explicitly.

C. *Consumer-Workers*

The typical consumer-worker in this moddel makes a sequence of choices. First, he chooses a gamble in wealth and realizes its outcome. Then he selects an occupation and, in so doing, adds the wage to the wealth already owned. Given the augmented wealth and occupation, he purchases insurance for the risk in the occupation. Finally, after realizing the outcome of the occupational risk and possibly receiving insurance payments, the consumer-worker experiences the satisfaction of the goods consumed.

The value of each choice in the sequence is dependent upon the choices that the consumer-worker will subsequently make. To characterize the sequence of optimum choices, one starts with the worker's ultimate utility and examines how that is optimized by appropriate insurance. The next step is to study choice of occupation, given that optimum insurance will subsequently be purchased. One then analyzes the choice of an optimum gamble, assuming that subsequent choices of occupation and insurance will be optimum.

1. *Choice of Occupation and Insurance*. Consumer-workers have identical utility functions of the form $U(z, x)$, where z is the indivisible good, taking the values of 0 for "dead" and 1 for "alive," and x is the divisible consumption good. Utility satisfies $U(1, x) > U(0, x)$ for all x, meaning that life is preferred to death, other things equal. Consumers are averse to gambles in x, given z, so that $U_2 > 0$ and $U_{22} < 0$. The indivisible good z enhances consumption of the divisible good x in the sense that $U_2(1, x) > U_2(0, x)$ for all values of x.

Consumer-workers of this sort desire to insure their risks of death. For a worker who has total wealth Y and probability of death equal to r', the expected utility with optimum insurance is

$$W(r', Y) = \text{Max}(1 - r')U(1, x) + r'U(0, y), \tag{10}$$

subject to

$$(1 - r')x + r'y = Y. \tag{11}$$

The constraint arises because workers purchase insurance at fair prices. A boundary solution of $y = 0$ is possible, but makes no fundamental changes in the results stated here.

Let $x = x(r', Y)$ and $y = y(r', Y)$ be solutions to this problem. From the first-order conditions of the consumer's maximization and from the envelope theorem,

$$W_2(r', Y) = U_2(1, x) = U_2(0, y), \tag{12}$$

$$W_1(r', Y) = -[U(1, x) - U(0, y)] + U_2(0, y)(x - y). \tag{13}$$

Now $W_{22} < 0$ follows from $U_{22} < 0$. From equation (12) and the enhancement property of life, one can readily infer that the optimum $x(r', Y)$, $y(r', Y)$ satisfy

$$x(r',Y) > Y > y(r',Y). \tag{14}$$

It can be shown that[8]

$$W_{21}(r',Y) < 0. \tag{15}$$

This means that the marginal utility of wealth is decreased by an increased chance of death. It rests ultimately upon the assumption that possession of the commodity "life" raises the marginal utility of consumption of the divisible good x.

Let I be the worker's wealth before wages. Then the consumer-worker chooses an industry so as to reach

$$T(I; w_s, w_r) = \mathrm{Max}[W(s+t, I+w_s), \\ W(r+t, I+w_r)]. \tag{16}$$

This $T(I; w_s, w_r)$ is the indirect utility function that guides the consumer at the time he chooses an optimum gamble.

2. *Choosing the Optimum Gamble*. The optimum gamble is the solution q_a, q_b, I_a, I_b, to the problem

$$\mathrm{Max}\, q_a T(I_a; w_s, w_r) + q_b T(I_b; w_s, w_r), \tag{17}$$

subject to

$$q_a I_a + q_b I_b = f(l_s) - w_s l_s + g(l_r) - w_r l_r + D \tag{18}$$

and

$$q_a + q_b = 1. \tag{19}$$

Equation (18) is the wealth constraint relevant to this gamble. The right-hand side is I_0, the endowed wealth per worker. This wealth comes from two sources: the share of each worker in the profits of each industry is 100 percent since labor force is unity, and each worker shares (at 100 percent) in ownership of the economy's endowment of goods D.

3. *Properties of the Consumer-Worker's Optimum*. This maximization is illustrated in Figure 3. Here $w_r > w_s$ because otherwise no workers would choose the risky industry at any level of wealth. The line tangent to $T(I; w_s, w_r)$ in two places denotes the optimum gamble. Since the consumer-worker enters industry S in one outcome of the gamble and industry R in the other, the incomes I_a and I_b will be replaced by I_s and I_r without loss of generality. Tangency is justified in this case by the first-order conditions

$$W(s+t, I_s + w_s) - W(r+t, I_r + w_r) \\ = W_2(r+t, I_r + w_r)(I_s - I_r), \tag{20}$$

$$W_2(s+t, I_s + w_s) = W_2(r+t, I_r + w_r). \tag{21}$$

It follows from (12) and (21) that

$$U_2(1, x(s+t, I_s + w_s)) \\ = U_2(1, x(r+t, I_r + w_r)), \tag{22}$$

and consequently that

$$x(s+t, I_s + w_s) = x(r+t, I_r + w_r). \tag{23}$$

By a similar argument,

$$y(s+t, I_s + w_s) = y(r+t, I_r + w_r). \tag{24}$$

This means that consumptions (x) of surviving workers and bequests (y) of accident victims are identical across industries. This is natural since the worker cares only about life and consumption, not about the industries per se. Viewed *ex ante*, the marginal utility of consumption is the same if he survives, no matter which industry he joins. There are markets in insurance and in gambles that permit him to equate marginal utilities of consumption among all states in which he is alive and, in fact, across all states. Thus,

---

[8] The proof of this assertion is derived from the comparative statics of the system of first-order conditions.

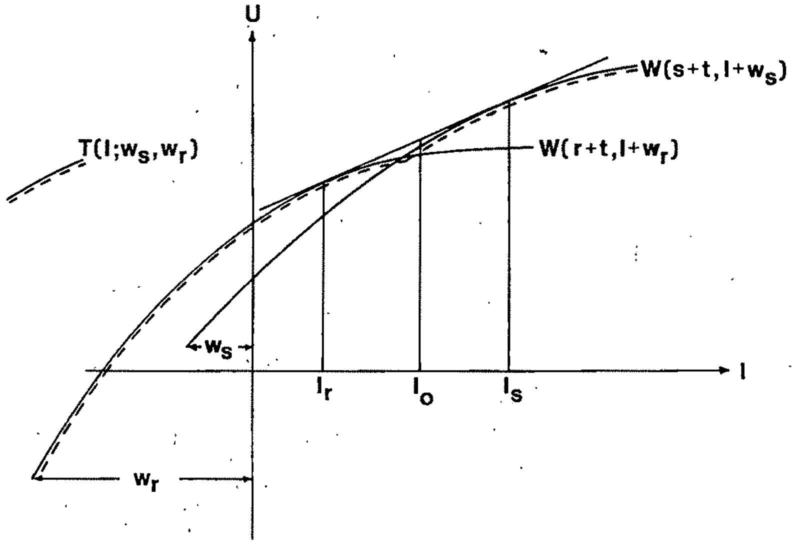


FIGURE 3. EQUILIBRIUM WAGES AND THE ASSOCIATED OPTIMUM GAMBLE

Note: $I_0 = q_s I_s + q_r I_r$.

consumptions contingent on life and death are the same in both industries. Since workers in R have higher risk of death, their expected utility $W(r+t, I_r + w_r)$ is less than the corresponding value for workers in S.

Equation (20) may be rewritten using the definitions of $W(r+t, Y)$ and $W(s+t, Y)$ and the relations $I_s = Y_s - w_s$ and $I_r = Y_r - w_r$ as

$$(25) \quad (r-s)[U(1,x)-U(0,y)] = U_2(0,y)(Y_s - w_s - (Y_r - w_r)).$$

Using the consumer's budget constraint, equation (11), to write Y_s and Y_r, one finds

$$(26) \quad Y_s - Y_r = (r-s)(x-y).$$

Using this relation, equation (25) can be rewritten in two useful forms,

$$(27) \quad \frac{[U(1,x)-U(0,y)] - U_2(0,y)(x-y)}{U_2(0,y)} = (w_r - w_s)/(r-s);$$

$$(28) \quad U(1,x) = U(0,y) + U_2(0,y)(x + (w_r - w_s)/(r-s) - y).$$

The left-hand side in the first of these forms is the marginal rate of substitution (MRS) between death-risk and consumption, $-W_1(r+t, Y_r)/W_2(r+t, Y_r)$ (see equations (12) and (13)). From equation (27), this MRS is the same in both industries and is equal to the wage difference divided by the probability difference. The left-hand side in equation (27) is referred to as the "value of life" in much of the literature, and the right-hand side is analogous to the estimates of this concept made by Richard Thaler and Sherwin Rosen (1976).

In equation (28), the quantity $(w_r - w_s)/(r-s)$ can be interpreted as the price of the indivisible good. This may be seen by comparing equation (28) to equation (7′) in Section I. The consumptions $I_a - p$ and I_b in equation (7′) are replaced by x and y that appear in equation (28), p is replaced by $(w_r - w_s)/(r-s)$, and $I_a - I_b$ is consequently replaced by $x + (w_r - w_s)/(r-s) - y$. Thus the optimum lottery in wealth followed by job choice produces the same consumptions and utilities as would be produced by an optimum lottery in wealth that would pertain if persons could subsequently purchase the indivisible good itself at price $(w_r - w_s)/(r-s)$.

D. *Equilibrium*

Equilibrium consists of quantities $q_s, q_r, I_s, I_r, Y_s, Y_r, l_s, l_r$, and prices w_s and w_r. The choices of quantities must be optimizing for firms and consumers given the prices, and market-clearing conditions in goods and labor must be met. The supply of labor to industry S consists of consumer-workers experiencing outcome I_s in their gambles, and similarly for industry R and I_r. For this reason, and because workers are a continuum in the model, the supply of labor to industry S is exactly q_s, and that to industry R is q_r. Thus the labor market clears when

$$q_s = l_s; \tag{29}$$

$$q_r = l_r. \tag{30}$$

The condition for clearing the commodity market is

$$q_s Y_s + q_r Y_r = f(l_s) + g(l_r) + D. \tag{31}$$

Market-clearing conditions for gambles and insurance are not explicit because of the assumptions made about supply in those markets.

Existence of this equilibrium is proved in the following section. Its character may be described briefly here. The market-clearing conditions permit us to substitute q_s and q_r for l_s and l_r wherever the latter appear. Profit-maximizing conditions imply that w_s and w_r can be replaced by $f'(q_s)$ and $g'(q_r)$. Because the marginal product of labor is very high at low labor inputs, it is inevitable that some workers are hired in each industry.

E. *Existence of Equilibrium*

The competitive equilibrium defined above solves a particular welfare-maximization problem. The welfare function is the mean of *ex ante* utility. This welfare function is not necessarily the appropriate one for any particular application. It is used here merely to show existence and to characterize equilibrium.

Let $Q(D,t)$ denote the solution to the problem

$$\text{Max}[q_s(1-s)+q_r(1-r)-t]U(1,x) + [q_s s + q_r r + t]U(0,y), \tag{32}$$

subject to

$$1 - q_s - q_r \geq 0, \tag{33}$$

$$D + f(q_s) + g(q_r) - [q_s(1-s) + q_r(1-r) - t]x - [q_s s + q_r r + t]y \geq 0. \tag{34}$$

Here the q_s and q_r are proportions of population employed in industries S and R. Constraint (33) restricts the supply of labor to that available—one unit. In equations (32) and (34), the expressions in brackets are the probabilities of survival and of death, respectively. In equation (34), $[q_s(1-s)+q_r(1-r)-t]x$ is the proportion of survivors multiplied by the consumption per survivor, and $[q_s s + q_r r + t]y$ is the proportion of casualties multiplied by the bequests per casualty. Constraint (34) means that total output equals or exceeds total consumption. The maximand is the average or expected utility.

Existence is demonstrated by rewriting the optimization problem. Eliminating q_s by use of constraint (33), write $v(q_r, Y) =$

$$(1 - s - q_r(r-s) - t)U(1,x) + (s + q_r(r-s) + t)U(0,y) \tag{35}$$

subject to

$$Y - (1 - s - q_r(r-s) - t)x - (s + q_r(r-s) + t)y \geq 0. \tag{36}$$

The first-order conditions for the maximization of this concave program include

$$v_Y = U_2(1,x) = U_2(0,y), \tag{37}$$

and from the envelope theorem, one may

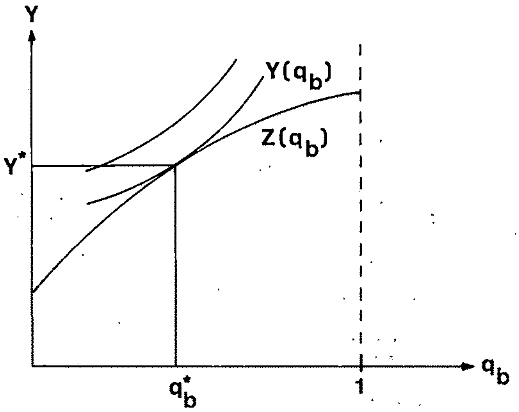


FIGURE 4. EXISTENCE OF EQUILIBRIUM AS A CONSTRAINED MAXIMUM

write

$$v_1 = -(r-s)[U(1,x)-U(0,y)] - U_2(0,y)(r-s)(x-y). \tag{38}$$

Consider the function thus defined as an objective function. An indifference curve of $v(q_r, Y)$ appears in Figure 4, denoted $Y(q_r)$. The slope $Y'(q_r)$ is the marginal rate of substitution between output and the probability of working in the dangerous industry. From $v_1 < 0$ and $v_2 > 0$, it follows that the indifference curve is upward sloping. It can also be shown that $Y''(q_r) > 0$. The maximization of social welfare is now rewritten

$$\text{Max}\, v(q_r, Y), \tag{39}$$

subject to

$$D + f(1-q_r) + g(q_r) - Y \geq 0. \tag{40}$$

The boundary of the constraint set is a concave function. It appears in Figure 4. Its slope is the marginal rate of transformation between total output and the labor input of the dangerous industry.

The optimum is characterized by equality of the marginal rate of substitution and the marginal rate of transformation. The curvature of indifference curves and constraints assures the existence of the optimum.

Now the task is to check that the optimum is an equilibrium. Obviously the market for labor clears when labor inputs are $l_s = q_s$ and $l_r = q_r$. The goods market clears because in equation (31), the left-hand side may be written $q_s[(l-s-t)x+(s+t)y]+q_r[(l-r-t)x+(r+t)y]$ and then (31) is the same as constraint (34) in the welfare maximization. Firms maximize profits when the wage rates are $w_s = f'(q_s)$ and $w_r = g'(q_r)$. It remains to check that at these wage rates the consumer maximizes expected utility by taking an optimum gamble with probabilities q_s and q_r and with prizes $Y_s - f'(q_s)$ and $Y_r - g'(q_r)$. However, the first-order conditions for the consumer's maximization in equations (17)–(19) are imitated by the first-order conditions in the maximization of $v(Y, q_r)$ at equations (39) and (40).

In particular, the condition in equation (20) that identifies the tangency of a single line to the two lobes of the indirect utility function is readily derived from the maximization of $v(Y, q_r)$. Among the first-order conditions for the latter maximization is

$$v_1 + v_Y(g'-f') = 0. \tag{41}$$

Substitute for the partials of $v(q_r, Y)$ from equations (37) and (38) and rearrange to yield

$$U(1,x) = U(0,y) + U_2(0,y) \times (x + (g'-f')/(r-s) - Y). \tag{42}$$

This equation imitates the necessary condition for utility maximization in equation (28).

The condition for clearing the commodity market (equation (31)) may be rewritten using $Y_s = I_s + f'(q_s)$ and $Y_r = I_r + g'(q_r)$, as

$$q_s(I_s + f'(q_s)) + q_r(I_r + g'(q_r)) = f(q_s) + g(q_r) + D. \tag{43}$$

But this is a rearrangement of the constraint on the optimum gamble, equation (18). That completes the required checks.

F. *Interpretation*

In equilibrium, the optimum gambles create a pattern of specialization that did not exist in their absence. The losers in the lottery acquire a comparative advantage in taking risks. Lower wealth and lower utility accompany this comparative advantage, although from the *ex ante* viewpoint, wealth and utility are the same for all persons and there are no differences in attitudes towards taking risks. An observer of the equilibrium outcomes will see wage differences that do not compensate risks in the sense of equalizing realized utility among those who bear greater and those who bear lesser risks. In fact, workers in the risky industry have the same contingent consumptions as do workers in the safe industry.

The wage differential is not great enough to compensate workers in industry S to change industries, but it is more than enough to compensate workers in R for remaining there. In this sense, $g'-f'$ is an underestimate of the cost of the increased danger for workers in industry S and an overestimate of that cost for workers in industry R. Since the idea of compensation by wage differentials has been fundamental at least since the work of Adam Smith, the situation is surprising. Smith's reasoning was that workers would move from the less-satisfying occupation to the more-satisfying one whenever wage differences failed to compensate. That mechanism fails here because workers having different outcomes of the optimum gamble in wealth freely choose different occupations and are unwilling to switch.

Wealths realized by workers in different industries are also unequal, since from $W_{21} < 0$, it can be shown that

$$(44) \qquad I_s + f'(q_s) > I_r + g'(q_r).$$

This inference rests ultimately upon the assumption that the marginal utility of the divisible good is higher if the consumer is alive.

The analysis can be given another interpretation. Let utility be additive over time and let $W(r',Y)$ be utility per year at risk equal to r' and wealth equal to Y. The worker may maximize utility by dividing his career among industries. Because $W_{21} < 0$, the worker wishes to consume less during his periods in the risky industry than he consumes while in the safe industry. If he spends fractions of time q_s and q_r in two industries and consumes at the rates Y_s and Y_r during those fractions of time, then he will achieve an optimum of utility just as in the previous paragraphs. The analysis could thus be phrased as optimum division of career rather than optimum lotteries. The difficulties raised for the idea of compensating wage differences would be the same as those that are described here in terms of optimum lotteries.

There is an special case in which the optimum gambles collapse and the wage differences are compensating. That is the case of $W(r',Y)=Y-h(r')$ where $h'>0$. Such utility is separable and displays risk neutrality toward gambles in wealth for fixed risk of death. Obviously it does not satisfy $W_{21}<0$. Gambles for the indivisible good vanish as the curves in Figure 3 approach more and more nearly to straight lines, and equilibrium has

$$(45) \qquad g'(q_r)-f'(q_s)=h(r)-h(s).$$

This means that the wage difference exactly compensates the risk difference. It follows that $Y_s = Y_r$ and $W(s,Y_s)=W(r,Y_r)$.

G. *The Shadow Price of Death*

The analysis above does not agree with the interpretation that has been made in practice of wage differentials in more- and less-risky industries. It raises the question of whether the wage difference is useful for making public decisions about risks. This question can be given a precise answer in the case that the welfare measure is *ex ante* expected utility.

Consider a risk of death t shared by all workers in the two-industry model. The risk produces an income $D(t)$ (or, equivalently, reducing the risk has a cost $-D(t)$) where $D'(t)>0$. For concreteness, one may think of the risk t as caused by workers all driving through a particular road intersection and $-D(t)$ as the cost of reducing risk t by installing increasingly sophisticated traffic signals. The risk involved is public in the sense of being common to all workers. Assuming that maximization of *ex ante* expected utility is appropriate, the optimum level of the public risk may be derived by maximizing $Q(D(t), t)$ over t, where Q is defined in equations (32)–(34). Second-order conditions are satisfied if $D''(t) < 0$. The condition determining optimum t is

$$D'(t) = -Q_t/Q_Y. \tag{46}$$

The Q_t and Q_Y may be derived using the envelope theorem. Then the optimizing conditions may be rewritten as

$$D'(t) = \left([U(1,x) - U(0,y)] - U_2(0,y)(x-y)\right)/U_2(0,y). \tag{47}$$

The left-hand side is the marginal rate of transformation between the public risk of death and income, and the right-hand side is the corresponding marginal rate of substitution (see equations (12) and (13)). From equation (27), this may be rewritten

$$D'(t) = (w_r - w_s)/(r - s). \tag{48}$$

Studies like that of Thaler and Rosen estimate differences in wages analogous to $w_r - w_s$. The quotient $(w_r - w_s)/(r - s)$, which they refer to as the value of life, is useful in optimizing conditions like that of equation (48) provided that the assumptions used in the present model are approximately satisfied. In particular, fair insurance and fair lotteries in wealth must be supplied.

Thaler and Rosen cited the compensating difference in wages as their basis for estimating the value of life from wage data. However, it is not exactly true that the criterion of equation (48) would be optimum in an economy with compensating wage differentials. An argument can be given for the approximate correctness of the criterion. Figure 5 illustrates the situation. Workers in the safe industry are at S and workers in the risky industry are at R. The curve connecting the industries is an indifference curve of $W(r', Y)$.

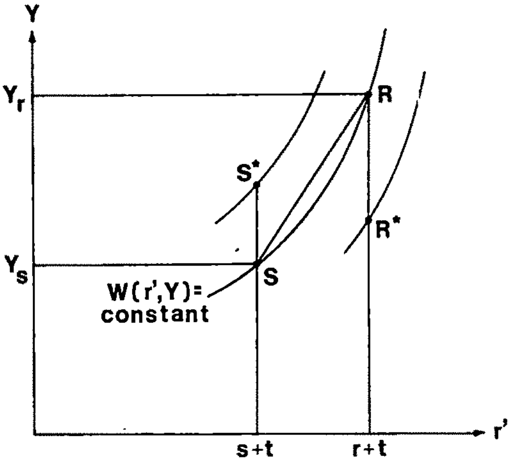


FIGURE 5. EQUILIBRIUM WITH (UNSTARRED) AND WITHOUT (STARRED) COMPENSATING WAGE DIFFERENCES

Income differences between industries arise solely from a difference in wage levels. This is so because there are no markets for gambles in wealth, such markets being incompatible with compensating differentials at equilibrium. It follows that $Y_r = I_0 + w_r$ and $Y_s = I_0 + w_s$ where I_0 is nonwage wealth. Thus, under the assumption of compensating wage differences, the difference in consumption $Y_r - Y_s$ is equal to $w_r - w_s$. Now the slope of the line SR, which is by definition $(Y_r - Y_s)/(r - s)$, must equal $(w_r - w_s)/(r - s)$. This slope is therefore the estimate of the value of life.

In this situation the risk is a Samuelsonian public good, and the criterion for optimum safety expenditure is to equate the marginal rate of transformation $D'(t)$ to a weighted sum of the marginal rates of substitution (*MRS*) at S and R. The *MRS*s are the slopes of the indifference curve at those points. The slope of line SR lies somewhere between the two *MRS*s, just as the weighted average does. As r approaches s, the *MRS*s converge toward each other, and the slope of line SR is squeezed between them. In the limit, $(w_r - w_s)/(r - s)$ approaches the required weighted average. In reality, the differences between safe and risky industries are rather great.[9]

---

[9] Thaler and Rosen discuss the magnitudes and conclude: "Thus, though the probabilities [of death in more hazardous occupations] are small in absolute terms, they are very large relative to the risks most people incur in the ordinary course of their lives" (p. 287).

So, under the assumption of compensating wage differences, the criterion $(w_r - w_s)/(r - s)$ remains approximate.

Figure 5 also illustrates the situation in an equilibrium that permits gambles in wealth. Workers in industries S and R are at S^* and R^*, respectively. Since the public risk is a Samuelsonian public good, the appropriate criterion is a weighted sum of *MRS*s at S^* and R^*. It is shown above that the marginal rates of substitution are the same at S^* and R^* and are in fact given by $(w_r - w_s)/(r - s)$. So the criterion of equation (48) is the one expected.

1. *Implications for Future Research.* The practical importance of these results is to place a burden of new empirical work upon students of the value of life. The burden arises because the Thaler-Rosen measure of the value of life is in theory dependent upon completeness of markets of a particular type. In order to sustain the theoretical argument supporting the Thaler-Rosen estimates, one needs evidence that the markets are effectively complete.

Casual empiricism seems not to confirm or deny the existence of the gambles discussed in this paper, perhaps because one is not accustomed to looking for such gambles. While gambling has been studied extensively, and indivisibility has frequently been cited as an incentive to gamble, there is no direct evidence that persons gamble in order to purchase indivisible goods. Something more than casual empiricism is therefore needed to support the Thaler-Rosen procedure.

To buttress the Thaler-Rosen method of estimating the value of life, it would be sufficient to show that the marginal rates of substitution between income and risk of death are approximately the same across industries. Showing this would imply either that the markets are effectively complete, or that indifference curves connecting income and risk are linear. In the latter case, there is no derived demand for gambles. In either case, one can rely on the Thaler-Rosen method of estimating the value of life.

Another approach would be to postulate that markets are incomplete and that the marginal rate of substitution between income and risk differs from one industry to another. Then one would try to argue that the Thaler-Rosen estimate is nevertheless approximately equal to the appropriately weighted average of marginal rates of substitution across industries. This argument is also empirical. The only means of sustaining it will be to produce estimates of the marginal rates of substitution within industries.

Estimates of marginal rates of substitution by industry may be hard to come by, but they will be of considerable practical use beyond their possible confirmation of the Thaler-Rosen estimates. In the presence of estimates that differ by industry, there would no longer be a reason to use the Thaler-Rosen method. Any particular public decision concerning risk will in practice affect some persons more than others. The decision on such public risks should take into account the preferences of the persons actually involved, and the estimates of the value of life by industry would supply such information.

III. Conclusion

The conclusions of this paper can be briefly summarized. Previous discussions of the value of life relied on the notion that wage differentials would compensate for differences in risk. However, if fair gambles in wealth are available, equilibrium wages do not compensate the worker for having a more dangerous job. The lack of compensating wage differences occurs because of the greater freedom permitted by the availability of gambles, not because of any constraint on the freedom to choose occupations.

The model of Section II assumes the existence of a market for gambles in wealth. The presence of this market makes the system of markets more complete than the one ordinarily considered. In the complete-markets model, the Thaler-Rosen estimate of the value of life is exactly the criterion that should be used in public safety decisions. In this sense, the model of Section II perfects and clarifies existing theory.

The improved theory raises questions about the Thaler-Rosen method of estimating the value of life. One way to answer these questions would be to try to confirm the existence of gambles of the kind required by the theory. In studying the possible existence of gambles, one should remember the tendency of the economy to create implicit markets for commodities that are not actually traded.[10] In a sense, the existence of a demand for gambles entails a mild presumption that the demand is somehow satisfied. If the explicit gambles in wealth suggested by the theory are in fact absent, gambles may be implicitly supplied by the operation of institutions that do not at first sight appear to involve gambles. Empirical studies aiming to confirm or deny the derived demand for gambles might concentrate on explicit gambling. Alternatively, they could examine institutions that supply gambles such as the military draft and the use of standardized tests for scholastic advancement.

The theories discussed in this paper are based upon the von Neumann-Morgenstern utility theory and upon the recognition of indivisibilities. It may be that the theory will prove to be empirically weak. If this is so, the reasons for it will be of interest: either the von Neumann-Morgenstern theory is at fault, the received theory of indivisibility is in error, or the costs of gambles are prohibitive. None of these alternatives seems probable. More likely, the gambles implied by the theory exist in forms that further research can reveal.

APPENDIX

Numerous examples of indivisible goods have appeared in the literature. In the paper by Paul Rubin and Chris Paul (1979), the indivisible good is the opportunity to reproduce. James Dowell's (1982) contribution regards the indivisible good as possession of a job. In the literature on urban rent, the indivisible good is the location of the dwelling. James Mirrlees (1972) states this fact clearly in his paper, and the welfare implications of the fact are discussed by David Levhari et al. (1978). Papers by Avinash Dixit (1973) and John Riley (1973), among others, involve the indivisibility of residential location. Theodore Bergstrom (1982) has examined the case in which the indivisible choice is between being a soldier or a civilian. Papers by J. S. Fleming (1969) and Nils Hakansson (1979) produce Friedman-Savage utility functions by introducing dislocations cost (Fleming) or borrowing constraints (Hakansson) into the model. These costs and constraints may be interpreted as stemming from indivisibilities.

There continues to be considerable discussion of whether the Friedman-Savage argument properly explains gambling or the simultaneous purchase of gambles and insurance. The present paper does not try to deal with the issues of that literature. The papers by Yew-Kwang Ng (1965), Fleming, Hakansson, and Dowell take up some of the issues.[11] Other papers appearing recently in the same area are by Benjamin Eden (1980), by Elie Appelbaum and Eliakim Katz (1981), and by Bailey et al. (1980).

---

[10] Rosen's study (1974) of product quality explores one aspect of this tendency, and the Thaler-Rosen investigation of the value of life explores another. While I have posed problems for the interpretation of the Thaler-Rosen estimates, the relevance of the implicit market in risk upon which the estimates are based is undoubtable.

[11] I am indebted to Dowell for giving me the opportunity to see his paper. This Appendix owes a good deal to it.

REFERENCES

Appelbaum, Elie and Katz, Eliakim, "Market Constraints as a Rationale for the Friedman-Savage Utility Function," *Journal of Political Economy*, August 1981, *89*, 819–25.

Bailey, Martin J., "Reducing Risks to Life: Measurement of the Benefits," American Enterprise Institute Studies, 1980.

______**,Olson, Mancur and Wonnacott, Paul,** "The Marginal Utility of Income Does Not Increase: Borrowing, Lending, and Friedman-Savage Gambles," *American Economic Review*, June 1980, *70*, 372–79.

Bergstrom, Theodore, "Soldiers of Fortune?," Discussion Paper, University of Michigan, May 1982.

Cook, Philip J. and Graham, Daniel A., "The Demand for Insurance and Protection: The Case of Irreplaceable Commodities," *Quarterly Journal of Economics*, February 1977, *91*, 143–56.

Dixit, Avinash, "The Optimum Factory Town," *Bell Journal of Economics*, Autumn 1973, *4*, 637–51.

Dowell, James, "Risk Preference and the Work-Leisure Trade-Off," manuscript, Monash University, 1982.

Eden, Benjamin, "The Insurance-Buying Gambler," *Economic Inquiry*, July 1980, *18*, 504–08.

Fleming, J. S., "The Utility of Wealth and the Utility of Windfalls," *Review of Economic Studies*, January 1969, *36*, 55–66.

Friedman, Milton and Savage, Leonard J., "The Utility Analysis of Choices Involving Risk," *Journal of Political Economy*, August 1948, *61*, 279–304.

Hakansson, Nils H., "Friedman-Savage Utility Functions Consistent with Risk Aversion," *Quarterly Journal of Economics*, August 1979, *84*, 472–87.

Jones-Lee, M. W., *The Value of Life: An Economic Analysis*, Chicago: University of Chicago Press, 1976.

Levhari, David, Oron, Yitzhak and Pines, David, "A Note on Unequal Treatment of Equals in an Urban Setting," *Journal of Urban Economics*, April 1978, *5*, 278–84.

Malinvaud, Edmond, "The Allocation of Individual Risks in Large Markets," *Journal of Economic Theory*, April 1972, *4*, 312–28.

Marshall, John M., "Moral Hazard," *American Economic Review*, December 1976, *66*, 880–90.

Mills, Edwin S. and MacKinnon, James, "Notes on the New Urban Economics," *Bell Journal of Economics*, Autumn 1973, *4*, 593–601.

Mirrlees, James A., "The Optimum Town," *Swedish Journal of Economics*, March 1972, *74*, 115–35.

Ng, Yew-Kwang, "Why do People Buy Lottery Tickets? Choices Involving Risk and the Indivisibility of Expenditure," *Journal of Political Economy*, October 1965, *73*, 530–35.

Riley, John G., "Gammaville—An Optimal Town," *Journal of Economic Theory*, October 1973, *6*, 471–82.

Rosen, Sherwin, "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition," *Journal of Political Economy*, January/February 1974, *82*, 34–55.

Rubin, Paul H. and Paul, Chris W., "An Evolutionary Model of Taste for Risk," *Economic Inquiry*, October 1979, *17*, 585–96.

Smith, Adam, *The Wealth of Nations*, New York: Random House, 1937.

Thaler, Richard and Rosen, Sherwin, "The Value of Saving a Life," in Nestor E. Terleckyj, ed., *Household Production and Consumption*, New York: National Bureau of Economic Research, 1976, 265–302.

Contractual Responses to the Common Pool: Prorationing of Crude Oil Production

By GARY D. LIBECAP AND STEVEN N. WIGGINS*

> Overproduction is like the weather, everyone talks about it, but no one has ever really done anything about it.
> [*Oil Weekly*, May 20, 1927]

There is growing interest in the role of bargaining costs in inhibiting voluntary contracts (Victor Goldberg, 1976; Oliver Williamson, 1979). The theoretical work stresses the problems of heterogeneous firms, asymmetrical information, and sequential adjustment as economic conditions change. There has, however, been little empirical work on the impact of the number and heterogeneity of parties on contracting success.[1] This paper addresses that issue by isolating firm differences that lead to differing bargaining positions, and by showing the high degree of concentration necessary to complete contracts. We analyze firm bargaining to mitigate rent dissipation from competitive production on five common oil pools where contracting success varied sharply. On some fields, agreements were quickly reached and effectively enforced, while on others compliance was never achieved. We show the firm heterogeneities that led to different bargaining positions, and calculate the level of concentration necessary for private contracting success. We measure concentration both by the inverse of the Herfindahl index and by the absolute number of firms.[2]

Three contractual solutions were considered to the common pool problem: lease consolidation, unitization of production under a single firm, and prorationing of field output among oil firms. In general, agreements for consolidation or unitization were not reached, and prorationing emerged as the dominant solution. Our analysis of prorationing contracts focuses on Oklahoma and Texas from 1926–35, a period when private and state controls were first implemented. The five largest fields discovered during that period are examined. The fields, their discovery dates, and 1949 cumulative U.S. production rankings were East Texas (1930, 1st); Oklahoma City (1928, 4th); Seminole District, OK (1926, 5th); Yates, TX (1926, 14th); and Hendrick, TX (1926, 22nd).[3] On Yates, consensus for prorationing was reached early, and dissipation margins were sequentially closed as they appeared. On Oklahoma City, Seminole, and Hendrick, there were partial agreements, and firm shares of the fixed field total were based on the number of wells; hence, dissipation occurred as firms drilled additional wells, and agreements broke down after short periods. On East Texas, field output limits were only effective during military occupation of the field in 1931 and under NIRA controls in 1933. Where private con-

---

*Department of Economics, Texas A&M University, College Station, TX 77843-4228. We thank Raymond Battalio, James Griffin, Charles Maurice, and other colleagues at Texas A&M. Helpful remarks were also provided from workshops at MIT and the Cliometrics Conference, Iowa City. Special thanks go to Victor Goldberg for his insightful comments. Financial assistance from the National Science Foundation through grant number SES-8207826 and from the Center for Energy and Mineral Resources, Texas A&M University, is gratefully acknowledged.

[1] Available empirical evidence on related issues is in Ronald Johnson and Libecap, 1982; Vernon Smith and Arlington Williams, 1981a, b; and Mark Isaac and Charles Plott, 1981.

[2] These measures of concentration emphasize different aspects of concentration on these fields. The absolute number of firms gives the number of parties in an agreement; the inverse of the Herfindahl index shows the degree of large firm control, but is insensitive to large numbers of small firms. The inverse of the Herfindahl index is the numbers equivalent of equal-sized firms (Morris Adelman, 1969). For example, a Herfindahl index of .33 gives a number equivalent of three equal-sized firms.

[3] Seminole District includes Seminole City, Earlsboro, Bowlegs, and Little River. See American Petroleum Institute, 1951, pp. 156–71.

tracting failed, firms lobbied for state enforcement, but again firm heterogeneities molded the final outcome.

Other studies of prorationing have criticized the rules for encouraging costly overcapitalization, nonoptimal pumping rates, and small firm biases (Morris Adelman, 1964, pp. 104–05; Melvin de Chazeau and A. H. Kahn, 1959, pp. 150–66; and James McKie and S. L. McDonald, 1962, pp. 113–16). These treatments have focused solely on production efficiency. We show that prorationing quotas were chosen to bring all firms into broad agreement, rather than to achieve minimum physical production cost. To achieve consensus for control of the most costly sources of rent depletion, concessions permitting less extensive dissipation along other margins were made. The provisions of the negotiated contract, then, while prohibiting particular activities by firms, set the lines along which some rent dissipation could occur. While the costs of the regulatory arrangement were high, our examination explains these rules as optimizing responses to achieve at least some control of rent dissipation.

In developing our argument, we first describe the options considered by firms to mitigate common pool conditions to show why prorationing was chosen. Section II provides a brief theoretical discussion to derive implications for the empirical analysis in Section III.

I. The Common Pool Problem and Potential Solutions

In this section we outline the characteristics of common oil pools and their impact on contracting: Part A describes the nature of the common pool and the costs of competitive extraction; Part B shows the link between fragmented surface ownership and the concentration of production, the key determinant of contracting success; and Part C examines the failure of consolidation and unitization.

A. *Characteristics of the Common Pool and Costs of Competitive Extraction*

Typically, oil reservoirs are compressed between an upper layer of natural gas and a lower layer of water. The two layers, as well as gas dissolved in the oil, drive the oil to the surface when the surrounding formation is punctured by a well. Oil migrates to the well, draining neighboring areas. The extent of migration depends upon subsurface pressures, oil viscosity, and the porosity of the rock. Reservoirs are not uniform; these characteristics differ across the field, generating inherent variation in well productivity. As a firm drills additional wells oil migrates more rapidly into the created low pressure zone, raising the firm's share of field output. Increases in the rate of production, however, reduce ultimate oil recovery. With high withdrawal rates, the ratio of natural gas and water to oil produced increases, leading to a greater loss in subsurface pressure. Pockets of oil became trapped, and retrievable only with high extraction costs.

Oil reservoirs are commonly found below numerous, independently owned surface tracts. The surface landowners initially hold the mineral rights, but transfer them to firms through mineral leases. By this process, multiple firms gain access to the pool. Under the rule of capture, property rights to crude oil are assigned only upon extraction (Robert Hardwicke, 1935). Given migratory oil and fragmented surface holdings, each producing firm has incentive to drill and drain. Rent dissipation follows with high capital costs—duplicate wells and surface storage—and with reduced total oil recovery. Discussions of unrestrained oil production in the 1920's and early 1930's emphasized extraordinary wastes (George Stocking, 1925; John Ise, 1926). The Federal Oil Conservation Board (1926, p. 30; 1929, p. 10) estimated recovery rates of only 20–25 percent with competitive extraction, while 85–90 percent was possible with controlled withdrawal. Excessive capital costs are indicated by the following data for the Hendrick field of West Texas. Discovered in June 1927, exploitation was essentially unrestricted until May 1928. During that time, competitive drilling by oil firms led to one well per 10 acres at a cost of $57,000 a well when, given porous geologic conditions, only one well per 80 acres would have rapidly drained the reservoir. The associated loss in underground pressure forced premature oil pumping at a marginal cost of

TABLE 1—CONCENTRATION OF LANDOWNERSHIP AND OPERATING FIRMS BY ACREAGE

	Numbers Equivalent		Absolute Number	
Field	Landownership	Operating Firms	Landownership	Operating Firms
Yates	1.3	3.2	2	16
Hendrick	2.1	7.5	3	18
Seminole[a]	120.0	15.0	127	40
Oklahoma City	20.0	2.0	42	7

Source: Oil Weekly, Oil and Gas Journal
[a] Seminole City pool

$.10 per barrel. Moderate withdrawal would have maintained pressure and allowed oil to flow without pumping until late in the field's life (*Oil Weekly*, March 23, 1928; April 13, 1928). Competition for rents necessitated large surface storage. During the first five months of 1928, storage capacity on Hendrick rose from 5,251,000 barrels to 10,987,000 barrels at a cost of $3,842,300. By contrast, on the neighboring Yates field where private controls limited production, storage was only 783,000 barrels and cost $274,000. This was despite a larger potential and greater ultimate oil recovery on Yates (*Oil Weekly*, February 24, 1928; March 23, 1928; May 25, 1928).

B. *Fragmented Land Ownership and the Concentration of Production*

The rent dissipation associated with common oil pools provides firms with clear incentives to restrict output. As we show, the number and heterogeneity of firms, however, limited the agreements that could be reached. The concentration of production was largely determined by surface landownership. In general, more fragmented surface holdings led to a lower concentration of operating firms on a field due to the leasing strategies of landowners and firms, an issue that is generally beyond the focus of this paper.[4] An important characterization of this process, however, linking surface holdings and firm concentration, is that large landowners fear drainage to neighboring areas, while small landowners seek oil migration to their lands. Accordingly, all landowners prefer *not* to lease to firms operating on adjacent tracts. Landowners also protect themselves by using short-term leases that expire if production does not occur rapidly. Additional incentive to drill is provided by granting firms up to 90 percent of the oil produced. The impact of fragmented land holdings on the concentration of production is shown in Table 1, where the absolute number and numbers equivalent of landowners and operating firms for Yates, Hendrick, Seminole, and Oklahoma City are presented. The concentration measures are calculated by acreage for the first date when such data are available. The data reveal that lower concentration of surface ownership generally leads to lower concentration of operating firms.[5] As we show below, the relationship described by the data in Table 1 is not sensitive to the stage of field development.

C. *Information Costs and Contractual Solutions to the Common Pool*

Three contractual possibilities are available to firms to mitigate rent dissipation: consolidation of production rights through purchase, unitization, and prorationing. Under unitization, production rights to the field are delegated to one operator with costs and revenues allocated among the firms based on their contribution to the unit. Consolidation and unitization assign production to a single firm, which then has the incentive to maximize field rents. Such agreements, however, are difficult to achieve because relative or absolute tract values for each firm must be determined, before production rights can be transferred. This requires estimates of oil in place, recoverability, and drainage. Given inherently high information costs, the importance of oil migration for small tracts, and variation in field characteristics, these estimates are costly and subject to large error.[6] Hence, it is difficult to estimate tract values and to determine whether the contract would make all parties better off. These problems limited unitization and consolidation, as shown by the following empirical data. By 1947, only 12 of some 3,000 fields in the United States were fully unitized (Joe Bain, 1947, p. 29).[7] For consolidation, Table 2 reports the numbers equivalent measures of concentration across time for the five fields in our sample. The dates selected in the table are based on available data. The data show that consolidation was not a general response to common pool conditions.[8] Indeed, the data reveal that the fields (except for East Texas) became less concentrated over time as new firms began producing on previously undeveloped land. For example, while there were 6 firms on the Yates field in September 1927, by July 1928 there were 20 (*Oil Weekly*, September 9, 1927; July 6, 1928).

---

[4] This phenomena is observed on maps of the five fields in the *Oil Weekly* and *Oil and Gas Journal*, that show both surface ownership and operating firms. See *Oil and Gas Journal*, August 4, 1927, for the Seminole City field. Leonard Logan (1930, pp. 52–54) discusses leasing strategies.

[5] Oklahoma City is an exception because the Indian Territory Illuminating Oil Company (ITIO) acquired production rights to the area before there was indication of productive reserves (*Oil Weekly*, December 14, 1928).

TABLE 2—CONCENTRATION OF PRODUCTION OVER TIME

Field	Numbers Equivalent	
Yates	1.9 September 1927	5.9 July 1928
Hendrick	10.0 September 1927	13.2 July 1928
Seminole City	11.2 May 1929	15.7 August 1929
Oklahoma City	5.0 December 1929	6.3 December 1932
East Texas	64.6 March 1931	35.3 February 1933

Source: Oil Weekly.
Note: Numbers equivalents for all fields except Seminole City are calculated by well ownership. For Seminole City, continuous well data were unavailable and output data were used.

The third option, prorationing, is a less complete but more flexible solution, which allowed agreements to be reached. Under prorationing, all parties are maintained on the field, and their output is controlled through simple quota assignments. Accordingly, once and for all tract values need not be estimated since no transfer of production rights occurs. Shares are automatically adjusted as production continues; some leases are flooded by water, and production declines on others due to pressure loss. Such adjustments are difficult to achieve with unitization and consolidation because both alter the pattern of development from what would occur with multiple producers. Prorationing allocation rules will be chosen to achieve consensus among the multiple firms exploiting the pool, and not all margins for dissipation will be controlled. The types of rent dissipation are determined by the quota arrangements. These arguments briefly explain why prorationing emerged instead of tract consolidation or field-wide unitization. The rest of the paper analyzes the determinants of the quota rules and impact of particular prorationing contracts.

II. Firm Size and the Incentive to Restrict Output

To clearly develop testable implications for the empirical analysis, we briefly summarize the common pool problem when firm sizes differ. We apply a variation of the dominant firm model to a common property resource. This approach illustrates how firm size affects both the willingness to enter agreements to control output and preferences for particular quota arrangements. Firm i's profit function is

$$\pi_i = pq_i - C_i(q_i, q_{\bar{i}})q_i, \tag{1}$$

---

[6] For example, a recent study of federal leasing showed that second place bids often ran 50 percent lower than the winning bid (E. C. Capen et al., 1971).

[7] Elsewhere (1983a) we examine the contracting problems that limited unitization.

[8] For Table 2 and throughout the paper, except as noted, we calculate concentration on the basis of well ownership, since acreage or production data are not consistently reported in the *Oil Weekly* or *Oil and Gas Journal*.

where $p =$ the parametric market price, $q_i =$ firms i's output, $q_{\bar{i}} =$ output of other firms on the field, and $C_i(q_i, q_{\bar{i}}) =$ firm i's average cost function.[9]

Oil migration leads to cost interdependence. Differentiating (1), yields the noncollusive profit maximizing output:[10]

$$(2)\qquad p = C_i(q_i, q_{\bar{i}}) + (\partial C_i/\partial q_i)q_i$$

$$= C_i(q_i, q_{\bar{i}}) + (\partial C_i/\partial q_i)S_iQ,$$

where $Q = q_i + q_{\bar{i}}$, and $S_i = q_i/Q$. The firm optimum is where price equals marginal extraction cost, which includes both the direct cost of additional output and the increased cost of inframarginal production. As S_i gets arbitrarily small, holding Q constant, inframarginal cost effects are completely external, and the firm produces where $p = C_i(q_i, q_{\bar{i}})$. As a firm's share increases, however, there is incentive to restrict output below $p = C_i(q_i, q_{\bar{i}})$.[11] Additionally, as shares increase, firms internalize more of the cost increases from rival production.

$$(3)\qquad \partial C_i/\partial q_{\bar{i}} = (\partial C_i(q_i, q_{\bar{i}})/\partial q_{\bar{i}})S_iQ > 0.$$

The cross-unit cost effects from common pool production show why large firms seek to limit field output; however, if they unilaterally reduce production, small firms will expand output. Thus, as large firms cut production, they must prevent expansion by small firms. These efforts to restrict output are further complicated by oil drainage. Any firm's output depends both upon oil in place below its lease and drainage from neighboring areas. Because of limited oil in place, small operators necessarily rely more heavily on drainage.[12] Drainage increases substantially with higher output rates, and, accordingly, small firms are reluctant to enter agreements to restrict output, despite reduced aggregate production costs and potential increases in ultimate recovery.

These arguments demonstrate the importance of firm size in determining positions on output controls. To obtain the support of small firms in restraining production, large firms will offer small firms quota concessions. Of the available options, only per well quotas provide gains to small firms to entice them into agreements, since they can drill additional wells to increase their share of output. An alternative allocation rule is acreage, but this clearly favors large firms by eliminating drainage. A similar disadvantage exists for quotas based upon past output, which do not allow small firms to increase their share of field production under regulation.

A number of testable implications directly follows. First, greater concentration of operating firms on a field will lead to more rapid and complete agreements, since greater concentration of production implies more large and fewer small firms. Second, per well allocations will be used on unconcentrated fields, while on concentrated fields acreage quotas will be used. Third, per well quotas will lead small firms to drill more densely than large firms to increase their share of field production. Finally, small firms are more likely to violate prorationing quotas. The differing incentives of firms to support prorationing and the corresponding bargaining concessions mean that prorationing will bring only partial controls on rent dissipation. The extent of uncontrolled dissipation will vary systematically across fields with operator concentration. We now turn to the empirical examination of bargaining for prorationing on the Yates, Hendrick, Seminole, Oklahoma City, and East Texas fields.

---

[9] The model is a static one with the primary characterization, cost interdependence. For dynamic considerations, see Paul Davidson (1963). While there are price effects associated with unconstrained output on the sample fields, those effects do not change the arguments presented here. For discussion see our (1983b) paper.

[10] We assume that firms follow a Cournot decision rule under conditions of free entry.

[11] Nonetheless, large firms will have lower direct production cost ($C(q_i, q_{\bar{i}})$) for any level of output ($q_i^*, q_{\bar{i}}^*$). With more acreage and given that oil is not perfectly migratory, each well draws upon a larger effective pool, implying lower cost at any output level.

[12] This technological effect is repeatedly described in the literature. For example, see Joseph Pogue, 1921, p. 32; *Oil Weekly* April 20, 1928; *Oil and Gas Journal*, July 28, 1927.

TABLE 3—CONCENTRATION AND CONTRACTING SUCCESS ON FIVE OKLAHOMA AND TEXAS OIL FIELDS

	Yates	Oklahoma City	Seminole	Hendrick	East Texas
Discovery Date:	July 1927	December 1928	July 1926	June 1927	October 1930
Numbers Equivalent of Firms Based on Well Ownership:[a]	1.9 (6) (September 1927)	5.0 (18) (December 1929)	14.0 (27) (May 1927)	10.0 (18) (September 1927)	64.6 (147) (March 1931)
Time from Discovery to Private Contract Agreement:	2 months	1 month	none completed	none completed	none completed
Time from Discovery to State Regulation:		1 year	1 year	10 months	7 months
Effectiveness of Output Controls:	Full compliance under private agreement	Early compliance, with small lot drilling violations increased and state regulations were necessary	Full compliance only with state control	Full compliance only with state control	No compliance except during military occupation in 1931 and the NIRA in 1933
Primary Allocation Rule:	Acreage under private agreement	Under state regulation, per well	Per well[b]	50% acreage 50% per well[b]	Per well

Source: Oil Weekly, selected issues.

[a] n, absolute number of firms shown in parentheses.

[b] The Hendrick and Seminole allocations were based on production potential, which could only be increased by drilling additional wells.

III. Contracting for Production Controls

A. *Field Concentration and Contracting Success*

Table 3 outlines the concentration of production and contracting success for the five fields. The concentration entries in the table, measured by the numbers equivalent, $1/H$, and the absolute number of firms, n, are for the first period when well ownership data are available. The table reveals a close relationship between the degree of concentration and contracting success. On the most concentrated field, Yates ($1/H=1.9$; $n=6$), agreements on well spacing, total production, and firm quotas were completed within two months of discovery. The initial quotas, based on a per well formula, soon led to overdrilling. The quota arrangement was then changed to productive acreage, and the margin was closed (*Oil Weekly*, September 9, 1927; December 2, 1927). An umpire was hired by the operators to monitor compliance and to arbitrate disputes. After the Texas Railroad Commission intervened in neighboring Hendrick, Yates operators requested state enforcement of their contract, a request based on fear of antitrust prosecution of the private agreement rather than its failure (*Oil Weekly*, June 15, 1928). At the initial agreement, there were 6 firms with completed wells, and one, the Mid Kansas Oil Company, had 71 percent of the field's 17 wells. A year later, Mid Kansas remained the dominant producer with 35 percent of the 203 wells on the field, and concentration remained high with $1/H$ equal to 5.9. Under the Yates plan, daily output was held below 150,000 barrels, even though the field's daily potential reached 4,000,000 barrels in

October 1928.[13] That restraint was despite low drilling costs associated with the field's shallow 1,000 foot depth ($15,000 per well compared with $57,000 on neighboring Hendrick: *Oil Weekly*, November 11, 1927; March 23, 1928). Moreover, Yates had the smallest surface storage investment of the five fields.

Oklahoma City, the next most concentrated field, had the next most successful early private contracting record. Bargaining began in December 1928, when the Indian Territory Illuminating Oil Company (ITIO) held 67 percent of the acreage (*Oil Weekly*, December 14, 1928). By December 1929 the field was still concentrated ($1/H = 5.0$, $n = 10$), and ITIO had 35 percent of the completed wells (*Oil and Gas Journal*, December 5, 1929). The first agreement limited the number of wells. The field's depth, 6,400 feet, led to high drilling costs of $155,000 per well (*Oil Weekly*, January 10, 1930); hence, the incentive to restrict drilling. Further, wells were closed for fixed intervals to reduce production. The field began southeast of Oklahoma City, but continued exploration extended the field into the city. Much smaller surface tracts (town lots) led to rapid entry by small firms and declining concentration. Between December 1929 and February 1932, the number of firms rose by 57, and 45 of the new entrants had fewer than 5 wells each. By contrast, 6 of the 10 early producers had over 20 wells each, and the largest, ITIO, had 293 (*Oil Weekly*, December 5, 1929; February 12, 1932). As the number and diversity of firms on the field increased, a consensus on private output restrictions could not be maintained and the larger operators petitioned the Oklahoma Corporation Commission to restrict output and prorate production (*Oil and Gas Journal*, September 19, 1929).

A per well allocation rule was chosen by the state in the face of large numbers of small firms. As with Yates, the resort to output limits helped to restrain early Oklahoma City production and to limit storage (*Oil Weekly*, August 23, 1929). Daily output remained stable at under 200,000 barrels from December 1929 through June 1933, despite a potential of over 3,000,000 barrels (*Oil Weekly*, November 17, 1930). The entry of small firms led to increased violation of production quotas, and output rose, peaking at 324,000 barrels in August 1933 (*Oil Weekly*, February 6, 1933; March 13, 1933).

Seminole and Hendrick were less concentrated than either Yates or Oklahoma City and had similar concentration and contracting records ($1/H = 14.0$, $n = 27$ for Seminole and $1/H = 10.0$, $n = 18$ for Hendrick). Neither field achieved successful private prorationing restrictions. At Seminole's peak output of 514,000 barrels per day in July 1927, all firms were producing at capacity (*Oil and Gas Journal*, July 28, 1927). Moreover, near its peak, Seminole had a storage capacity of 3,000,000 barrels, over three times that of the larger Yates field (*Oil Weekly*, March 4, 1927). No constraints were implemented on Hendrick and Seminole until the Texas Railroad Commission and the Oklahoma Corporation Commission intervened at the operators' request. State prorationing began in the Hendrick field in May 1928. On Seminole, formal controls appear not to have been in place until August 1928, nearly a year after the first petition to the Corporation Commission (*Oil Weekly*, July 8, 1927; August 10, 1928). As implied by the theory, prorationing rules in both fields relied upon per well allocations.

Table 3 shows that East Texas had both the least concentration and the least success in restricting output through private or government controls. Discovered in October 1930, the shallow field (3,500 feet) had low drilling costs (approximately $26,000 per well) and diffuse surface ownership that led to rapid entry by small firms and competitive drilling (*Oil Weekly*, March 6, 1931). By March 1931, the numbers equivalent based on well ownership was 64.6.[14] By early 1933, there were over 1,000 firms and 10,000 wells on the field. East Texas production reached 738,035 barrels a day in August 1931 and peaked in May 1933 at 1,074,180 barrels. Rapid extraction reduced subsurface pressure, and lowered total recovery. Despite the costs of unrestrained production, no private prorationing contracts were completed. The Railroad Commission intervened with little success beginning in April 1931. The failure of these efforts led the Texas governor to close the field twice under martial law in August 1931 and in December 1932. State-imposed quotas were violated and numerous successful court challenges to government prorationing were initiated. Except during military occupation, the field remained uncontrolled until the NIRA codes were implemented and federal limits on interstate shipment of hot oil (oil produced above quotas) were imposed. This failure occurred despite exclusive use of a per well allocation rule by the Railroad Commission.

---

[13] *Oil Weekly*, October 12, 1928. Field potential estimates for all fields are overstated because they were based on wide-open production by each well for selected short periods. If all wells were run continually, pressures would have dropped and reduced field production.

[14] Because of rapid drilling on the field, the concentration was calculated on the basis of ownership of both completed wells and wells being drilled. During early development, there were almost no dry holes on East Texas.

The most important feature in the success or failure of contracting was the degree of concentration on each field; other factors were not systematically related with bargaining outcomes: the two fields with the lowest drilling costs, Yates ($15,000) and East Texas ($26,000), had opposite contracting results; Oklahoma City with drilling costs of $155,000 had early production agreements, while on Hendrick, where costs were $57,000 per well, controls were implemented relatively late and only with state support. Moreover, size of field was unrelated. Oklahoma City and East Texas were the largest fields in terms of production, but their contracting patterns were similar only when concentration declined in Oklahoma City with development on fragmented surface plots.

B. *Firm Size and Support for Prorationing*

The theory implies that large firms will support prorationing and small firms will resist. The concentration data support this implication, but we also have more detailed evidence of the position of firms on prorationing. Support for prorationing is evidenced by joining private agreements to restrict output, by complying with state prorationing rules, by promoting prorationing through testimony before regulatory agencies and legislatures, or membership in advisory bodies for implementing controls. Opposition to prorationing is evidenced by failure to comply with private or state restrictions, initiation of court challenges of prorationing, or testimony before regulatory agencies and legislatures against constraints.

On the Seminole field, there are published lists of the votes of firms on early prorationing controls. In May 1927, a meeting of 20 Seminole operators was held to restrain drilling. The 16 operators listed in favor held 73 percent of the wells on the field, and 12 of those were among the 15 largest firms by well ownership (*Oil Weekly*, May 20, 27, 1927). Barnsdall, the eighth largest firm, voted against the particular rule under discussion, but the firm was active in other prorationing efforts. The remaining three opposing votes were by firms with only 2 of the 447 completed wells on the field. A prorationing advisory committee of 5 firms was established in August 1927 to restrict exploratory wells; the members were all top 10 firms by well ownership (*Oil Weekly*, May 27, 1927; August 5, 1927). Later in 1928, 7 firms agreed to limit drilling on their leases. Six were ranked in the top 10 on the field (*Oil Weekly*, May 27, 1927; February 17, 1928). In 1929, 17 firms (of 34) voluntarily closed operations on Sundays to supplement formal controls; that included 7 of the largest 10 firms on Seminole (*Oil Weekly*, February 22, 1929; March 1, 1929; *Oil and Gas Journal*, January 19, 1928).

On the Oklahoma City field, the initial drilling restriction of December 1928 was promoted by the field's largest firm, ITIO, a leading advocate of prorationing. Once controls were in place, large firms voluntarily cut production below their authorized quotas. In July 1931, 365 wells, nearly half the wells of the field, were voluntarily closed to reduce output; 212 of those were owned by 5 of the largest 10 firms (*Oil Weekly*, July 24, 1931; February 12, 1932). By September 1931, voluntary underproduction by large firms totalled 3,600,000 barrels; 71 percent was by ITIO, Slick (2nd largest number of wells on the field), Phillips (3rd), Franklin (5th) and Skelly (8th) (*Oil Weekly*, September 25, 1931; February 12, 1932). The record for small operators is sharply different. Between September 1930 and September 1932, published sources list 16 firms, all small, as challenging or violating Oklahoma City prorationing rules. Of those, only 2 had as many as 10 wells, and the rest had 5 or fewer wells.[15] Through 1935, none of the 10 largest firms on the Oklahoma City field initiated court suits or was identified as producing beyond quotas.

In East Texas, the division between large and small firms was similar. When the Railroad Commission began to regulate the field in April 1931, an umpire was chosen to set and police quotas; costs were covered through firm assessments, but only large firms paid (Texas Senate *Journal*, 1931, pp. 15–23). In testimony before the Legislature, 4 of the largest 24 operators on the field (Gulf, Texas Company, Arkansas Fuel, and Sinclair) offered to support *any* prorationing bill, regardless of its quota structure (*Oil Weekly*, July 3, 1931). Humble and Gulf, the largest firms on East Texas, attempted private enforcement of prorationing rules by offering higher prices to firms that complied with quotas and by refusing to purchase oil produced in excess of quotas (*Oil Weekly*, April 24, 1931; May 29, 1931). By contrast, small firms opposed prorationing through lobby efforts in the Legislature, failure to comply with quota, and court suits. Their opposition to East Texas controls continued despite the use of per well quotas that allowed them to drill additional wells (*Oil Weekly*, July 3, 1931). Of more than 50 firms identified between 1931 and 1933 in court suits against Texas Railroad Commission's prorationing rules, or listed as violators, only one was among the top 24 firms on East Texas.[16] Small firm opposition is illustrated by one case involving 7 firms with combined holdings of only 94 acres, but daily production of 50,000 barrels. Their production of approximately 500 barrels per acre per day far exceeded the field average of 10 barrels per acre, and was due largely to drainage. They challenged Railroad Commission quotas that would have reduced their strategic advantage (*Oil Weekly*, July 31, 1931; June 5, 1931).

C. *Concessions on Quotas and their Impact on Behavior*

While small firms prefer per well quotas because they permit them to expand production, such quotas lead to the overdrilling documented by Adelman and others. The role of quota concessions is clearly shown in the negotiations on Hendrick, Yates, and East Texas. During initial contracting on Hendrick, there were more wells on tracts operated by small firms than on those held by large firms: the numbers equivalent based upon acreage was 7.4, while for wells it was 10.5. A major opponent of prorationing was Cranfill, a small firm with only 5 percent of the acreage, but 9 percent of the field's output (*Oil Weekly*, April 20, 1928). To reduce opposition by small firms, a quota was adopted based on 50 percent wells and 50 percent acreage. Six months after the quotas were installed, additional drilling by small firms reduced well ownership concentration to 13.3 (*Oil Weekly*, April 20, 1928; July 6, 1928; November 1, 1928). On Yates, the limited opposition to acreage quotas came from one small firm (*Oil Weekly*, December 9, 1927). The firm, Simms, had 9 percent of the field's production potential, because of wells located strategically for drainage, but only 6 percent of the acreage (*Oil Weekly*, December 23, 1927; January 27, 1928). Opposition by Simms did not lead to a breakdown of private controls, but when state regulation was instituted, a 50 percent well, 50 percent acreage quota was adopted. The change from acreage allocations immediately led to the drilling of 51 new wells at a cost of $750,000 (*Oil Weekly*, May 25, 1928). In East Texas, there were no effective controls until martial law was declared by the Texas governor in August 1931, because of opposition by small firms who held 85 percent of the leases on the field (Texas Senate *Journal*, 1931, p. 291).

---

[15] Violators of prorationing rules were listed in the *Oil and Gas Journal*, January 2, 9, 1931 and the *Oil Weekly*, September 12, 1930; January 22, 1932; September 12, 19, 1932; October 16, 1932. See also American Bar Association (1938, pp. 175–211).

[16] The *Oil Weekly* from April 1931 through December 1932 contains articles on court suits. Articles with lists of parties can be found in the issues of May 22, 1931; June 5, 1931; March 4, 1932 and May 16, 1932. Some lists of court injunctions against violators are reported in the *Oil Weekly*, January 9, 16, 23, 1932, and February 13, 1932.

We have additional evidence from East Texas that per well quotas lead small firms to drill more densely than large firms. The quota adopted by the Railroad Commission in September 1931 granted 225 barrels per day per well. Under that rule, small firms drilled additional wells and large firms held back. For the next year and a half, an average of 110 wells were sunk each week (*Oil Weekly*, January 30, 1933). By February 1933, the 24 largest firms on the East Texas field had an average of one well per 14 acres, while smaller firms averaged one well per 9 acres (*Oil Weekly*, February 27, 1933). The Cole Committee (U.S. House of Representatives, 1939, p. 503) reported that prorationing rules in East Texas contributed to the drilling of 23,000 unnecessary wells at a per well cost of $26,000.[17] As the number of wells rose, quotas were reduced to maintain targeted field output levels. They fell from 225 barrels per well in September 1931 to 37 barrels by December 1932 (J. H. Marshall and N. L. Meyers, 1933, p. 717). Texas law, however, gave special quotas of 40 barrels to high-cost pumping wells, more than the quota available to free-flowing wells in East Texas. Accordingly, firms placed pumping units on their wells at a cost of $3,500 per well to qualify for the larger allocation (*Oil Weekly*, January 23, 1933; February 27, 1933; *Railroad Commission of Texas v. Rowan and Nichols*, 107 Fed 2nd 70).

IV. Concluding Remarks

This paper has examined bargaining among firms to mitigate rent dissipation. Because of high bargaining costs, more complete solutions, consolidation and unitization, were not chosen. Instead, firms resorted to prorationing, and success varied. On Yates and Oklahoma City, where initial concentration of output was high, early agreements were reached. Private efforts failed on the less concentrated Seminole and Hendrick fields, though sufficient agreement was mustered to request state intervention to enforce compliance. On East Texas, with hundreds of small firms, no consensus could be reached on either private or government controls. Indeed, a major finding of the research is the high degree of concentration necessary to achieve successful coordination. Private agreements were completed when the numbers equivalent was less than 5; state enforcement brought compliance when it was approximately 10–12; but when concentration was less, state intervention could not control production without the use of troops. This is new empirical evidence on the contracting process among heterogeneous firms and the importance of concentration levels on the outcome.

The analysis of contracting also shows why prorationing took the form it did. Adelman and others have criticized prorationing for encouraging unnecessary wells and have called for oil field unitization. Adelman (1964, p. 107) estimated that prorationing raised production costs $4 billion annually. To require that regulatory policies fully close all margins for dissipation fails to recognize the high contracting costs for reaching agreement. We show that concessions, such as per well quotas, were required to draw in small operators and that the quotas led to predictable responses regarding rent dissipation. Nevertheless, given Adelman's figures, it is extraordinary how much waste was embodied in the regulatory outcome.

Despite those costs, prorationing controlled total field production and costs. Further, prorationing controls conserved natural reservoir energies and lengthened field life. On Seminole, state-enforced prorationing may have reduced early output by as much as 50 percent (*Oil Weekly*, February 17, 1930). Available data indicate that lower withdrawal rates increased ultimate recovery and lengthened the productive life of oil fields. The prorationing regulations examined in this paper followed the major oil discoveries of 1926–35, which increased both the magnitude of the technological losses from common pool production and the gains from contracting to limit output. When private agreements failed, the parties successfully appealed for state enforcement, and a permanent regulatory structure was established. Erich Zimmermann (1957, pp. 286–88) compared the productive life of 20 fields in Arkansas, Louisiana, Oklahoma, and Texas, 10 discovered prior to formal regulation and 10 discovered after regulation.[18] After 15 years, output from the earlier group had declined to an average of 8.6 percent of peak output, while the latter group produced an average of 73.9 percent of peak production. The sampled fields in this paper are transitional fields where production occurred under both uncontrolled and controlled conditions. For Yates, Hendrick, Seminole, Oklahoma City, and East Texas, fifteenth-year production was 28 percent of peak output.[19] In the late 1940's, industry sources cited by Zimmermann (p. 281) estimated that prorationing rules in Texas, Louisiana, New Mexico, and Mississippi annually saved the drilling of 100,000 unnecessary wells at an average per well cost of $100,000, a saving of roughly $10 billion. Thus, we see that prorationing was an effective partial solution to rent dissipation.

---

[17] Engineers testified before the Cole Committee that 3,000 wells could have drained the field. While interest rates and expected price changes influence optimal extraction, they would not explain the drilling of 23,000 additional wells. Clearly, competitive extraction was a driving force in the decision to drill.

Our examination of contracting has major implications for the general analysis of regulatory institutions. Without recognition of the heterogeneities among firms and their impact on contracting efforts to increase rents, the observed prorationing arrangements that emerged in Oklahoma and Texas cannot be explained. Similar heterogeneities influence regulations elsewhere in the economy. Accordingly, detailed analysis of bargaining among firms is essential for insight into the emergence of various institutional forms.

---

[18] Arkansas and Louisiana adopted formal regulation during the same period as Oklahoma and Texas (Stephen McDonald, 1971, p. 40).

[19] Calculations for the ratio of fifteenth-year (from discovery) production to peak production were based on field output data from the *Minerals Yearbook*, U.S. Department of the Interior.

REFERENCES

Adelman, Morris, "Efficiency of Resource Use in Crude Petroleum," *Southern Economic Journal*, October 1964, *31*, 101–22.

_____, "Comment on the '*H*' Concentration Measure as a Numbers Equivalent," *Review of Economics and Statistics*, February 1969, *51*, 99–101.

Bain, Joe, *The Economics of the Pacific Coast Petroleum Industry, Part III*, Berkeley: University of California Press, 1947.

Capen, E. C., Clapp, R. V. and Campbell, W. M., "Competitive Bidding in High Risk Situations," *Journal of Petroleum Technology*, June 1971, *23*, 641–53.

Davidson, Paul, "Public Policy Problems of the Domestic Crude Oil Industry," *American Economic Review*, March 1963, 53, 85–108.

De Chazeau, Melvin G. and Kahn, A. H., *Integration and Competition in the Petroleum Industry*, Petroleum Monograph Series, Vol. 3, New Haven: Yale University Press, 1959.

Goldberg, Victor P., "Regulation and Administered Contracts," *Bell Journal of Economics*, Autumn 1976, *7*, 426–52.

Hardwicke, Robert E., "The Rule of Capture and Its Implications as Applied to Oil and Gas," *Texas Law Review*, June 1935, *8*, 391–422.

Isaac, Mark and Plott, Charles, "The Opportunity for Conspiracy in Restraint of Trade: An Experimental Study," *Journal of Economic Behavior and Organization*, March 1981, *2*, 1–30.

Ise, John, "*The United States Oil Policy*, New Haven: Yale University Press, 1926.

Johnson, Ronald N. and Libecap, Gary D., "Contracting Problems and Regulation: The Case of the Fishery," *American Economic Review*, December 1982, *72*, 1005–22.

Logan, Leonard M., Jr., *Stabilization of the Petroleum Industry*, Norman: University of Oklahoma Press, 1930.

McDonald, Stephen L., *Petroleum Conservation in the United States*, Baltimore: John Hopkins University Press, 1971.

McKie, James W. and McDonald, S. L., "Petroleum Conservation in Theory and Practice," *Quarterly Journal of Economics*, February 1962, *76*, 98–121.

Marshall, J. H. and Meyers, N. L., "Legal Planning of Petroleum Production: Two Years of Proration," *Yale Law Journal*, March 1933, *42*, 702–46.

Pogue, Joseph E., *The Economics of Petroleum*, New York: Wiley & Sons, 1921.

Smith, Vernon and Williams, Arlington, (1981a) "The Boundaries of Competitive Price Theory: Convergence, Expectations and Transactions Costs," paper presented at Public Choice Meetings, New Orleans, March 1981.

______ **and** ______, (1981b) "Cyclical Double Auction Markets with and without Speculators," paper presented at Southern Economic Association Meetings, New Orleans, November 1981.

Stocking, George W., *The Oil Industry and the Competitive System, A Study In Waste*, Boston: Houghton Mifflin, 1925.

Wiggins, Steven N. and Libecap, Gary D., (1983a) "Bargaining Costs and Unitization of Oil Fields," Department of Economics, Texas A&M University, 1983.

______ **and** ______, (1983b) "Contracting by Heterogeneous Firms and Government Regulation of Crude Oil Production," Department of Economics, Texas A&M University, 1983.

Williamson, Oliver E., "Transactions Costs Economics: The Governance of Contractual Relations," *Journal of Law and Economics*, October 1979, *22*, 233–61.

Zimmermann, Erich W., *Conservation in the Production of Petroleum*, New Haven: Yale University Press, 1957.

American Bar Association, *Legal History of Conservation of Oil and Gas*, Chicago, 1938.

American Petroleum Institute, *Petroleum Facts and Figures 1950*, New York, 1951.

Federal Oil Conservation Board, *Complete Record of Public Hearings*, Washington: USGPO, 1926.

______, *Report III*, Washington: USGPO, 1929.

Oil and Gas Journal, various issues.

Oil Weekly, various issues.

Railroad Commission of Texas v. Rowan and Nichols, 107 Fed 2nd 70.

Texas Senate *Journal*, First Called Session, 42nd Legislature, Austin 1931.

U.S. Department of the Interior, *Minerals Yearbook*, Washington: USGPO, 1943–47.

U.S. House of Representatives, Cole Committee, "Hearing on House Resolution 290 and HR 7372 to Promote the Conservation of Petroleum; to Provide for Cooperation with the States in Preventing Waste of Petroleum; to Create an Office of Petroleum Conservation," 76th Congress, Washington, 1939.

Price Expectations of Business Firms: Bias in the Short and Long Run

By FRANK DE LEEUW AND MICHAEL J. MCKELVEY*

The results of surveys of price expectations by households and businesses do not provide much support for rational (i.e., unbiased and efficient) expectations. Edward Gramlich (1983) finds household price expectations to be biased, with serially correlated errors (a sign of inefficiency). Jonathan Leonard (1980) finds wage expectations for large companies to be biased and inefficient, and our own 1981 study finds business price expectations to be biased, with errors related to widely known macroeconomic variables (another sign of inefficiency).[1] Critics have suggested that these findings may be due not to any properties of "true" price expectations, but rather to errors in reporting expectations in business and household surveys.

More noteworthy than this evidence of possible bias or inefficiency is another implication of these studies; the finding that a 1 percent increase in the actual rate of price change, even if it is maintained for a very long time, leads to an increase of less than 1 percent in the expected rate of price change. We shall refer to this finding as the long-run bias proposition. If true, it undermines not only rational expectations views, but also any theories that rule out long-run "fooling" of market participants with respect to the rate of inflation.[2]

The principal goal of this paper is to subject the long-run bias proposition to close scrutiny, using price expectations data from the Bureau of Economic Analysis (BEA) Plant and Equipment Survey. In contrast to our previous paper using BEA data, which relied solely on weighted aggregates for two-digit industries, this paper makes use of a file containing individual company data for five years. There is much more scope in this paper than in the earlier one for examining the data intensively and disaggregating it in alternative ways. As we shall see, this greater scope leads us to reject the long-run bias proposition. The BEA data, after some adjustments, are consistent with an eventual point-for-point response of inflation expectations to actual inflation.

An important key to the new results is a five-way disaggregation of firms into (a) firms in the "regulated" (utility and communications) industries, (b) large firms in manufacturing, mining and trade, (c) medium-sized firms in manufacturing, mining and trade, (d) small firms in manufacturing, mining and trade, and (e) firms in the construction, services, and transportation industries. The first two categories—regulated firms and large firms—behave quite differently from the other three. The first two make quite accurate forecasts of their sales prices, but are no more accurate than other firms in forecasting the prices of capital goods they purchase. The first two depart in major respects from an error-learning hypothesis about the formation of sales price expectations; but that hypothesis, especially when augmented by an additional price-accelera-

---

*Bureau of Economic Analysis, U.S. Department of Commerce, Washington, D.C. 20230. Views expressed are our own. We appreciate helpful comments by Paul Wachtel, Robert Litterman, and James O'Brien on an earlier draft.

[1] Studies of price expectations by professional forecasters, in contrast to expectations by households and businesses, show results from which different analysts have drawn different conclusions. Expectations by forecasters, however, do not influence behavior of market participants unless they affect the participants' own expectations. In this study, we do not concern ourselves with professional forecasters' expectations.

[2] Both Gramlich and our 1981 study have this implication if the actual rate of inflation is raised by 1 percent and all other variables are held constant. In Gramlich's study, the same result holds even if not only the actual rate of inflation but also the growth of the money supply is raised by 1 percent. In our earlier study, the actual inflation-plus-money growth result holds for sales prices, but not for capital goods prices.

tion term, accords well with capital goods price expectations in these categories, and with both sales and capital goods price expectations in the other categories. Our conjecture is that the reason for these differences is that firms in the first two categories have much more market power in setting their sales prices than do other firms, and that when they report sales price expectations, they are reporting *plans* rather than *forecasts*. For the price expectations that we believe represent forecasts, an augmented error-learning model suggests very little long-run bias.

The paper is organized into five sections following this introduction. The first is a section on data. Next is a comparison of regression results for individual firms with results for groups of firms, a comparison which leads us to stress grouped data in the remainder of the paper. Then follows a section analyzing short-run bias in the grouped data, utilizing the five-way disaggregation described above. The fourth section is a regression analysis of the formation of price expectations, focusing on the long-run bias proposition. The final section draws some conclusions.

I. Data

The data are responses to questions in the year-end surveys of business expenditures on plant and equipment conducted by BEA since 1970. A question about price expectations is asked separately for prices of capital goods purchased and prices of goods and services sold. The question in the survey conducted at the end of 1980, essentially the same as in earlier years, was: "What are your best estimates of average price changes from 1979 to 1980 and from 1980 to 1981? (approximations are acceptable)." The survey included a box for entering perceptions of the past (1979–80) price change and one for entering expectations for the future (1980–81) price change (both in percent). We restricted the observations in our files to firms answering for at least two consecutive years, in order to test the influence of lagged variables. With this restriction, we were able to analyze data for 3,000–3,600 firms for each year from 1976 through 1980. In all, we had 15,742 observations for capital goods prices and 17,852 observations for sales prices.

The fact that respondents to the BEA survey report on actual price changes as well as expected price changes is an advantage of this data set over some others; for example, over the household price expectations series derived from the Michigan Survey (see Thomas Juster and Robert Comment, 1980). Elsewhere, we (1982) have analyzed the relationship of actual prices reported in the BEA survey to official price indexes; in brief, correlations of year-to-year price changes by two-digit industry are generally high, but there are some statistically significant differences.

Some characteristics of the data appear in Table 1. By industry, a little over one-third of the observations are in manufacturing and a little over one-quarter in services. By size, where a size classification is available,[3] medium-size firms outnumber large firms by about three to one, and small firms are represented still more heavily. By year, the division is fairly even among the five years included. By reported price change, there is a concentration in the 4–10 percent range for sales prices and the 6–15 percent range for capital goods prices. Sales price changes are much more dispersed than capital goods price changes.

II. Grouped vs. Ungrouped Data

As soon as we begin to scrutinize data for individual firms, the possibility of errors of measurement strongly suggests itself. Consider, for example, the following example, by no means the only one of its kind, of capital goods price changes reported and expected by one firm (shown in percent):

	Reported change	Expected change
1976	15	18
1977	15	18
1978	10	15
1979	15	15
1980	15	20

---

[3] In the BEA survey, only firms in manufacturing, mining, and trade are classified into several size groups. Firms in other industries are classified into only two groups: selected very large firms and all other.

TABLE 1—DATA CHARACTERISTICS, BEA INDIVIDUAL COMPANY PRICE REPORTS

	Sales Prices		Capital Goods Prices	
	Number	Percent	Number	Percent
Total	17,852	100	15,742	100
Industry:				
Mining	399	2	408	3
Manufacturing	6,073	34	5,841	37
Trade	2,718	15	2,009	13
Transportation	2,077	12	1,491	9
Utilities and Communications	693	4	838	5
Construction	925	5	754	5
Services	4,967	28	4,401	28
Size:				
Manufacturing, Mining, Trade, and Selected Services:				
Large	1,495	9	1,740	11
Medium	5,256	29	5,059	32
Small	7,625	43	5,387	34
All other[a]	3,476	19	3,556	23
Year:				
1976	3,926	22	3,453	22
1977	3,555	20	3,199	20
1978	3,356	19	3,062	20
1979	3,336	19	2,975	19
1980	3,679	21	3,053	19
Reported Price Change (percent):				
Less than zero	1,513	8	142	1
0.1 to 2.0	533	3	86	1
2.1 to 4.0	1,003	6	147	1
4.1 to 6.0	3,971	22	1,768	11
6.1 to 8.0	3,803	21	3,182	20
8.1 to 10.0	3,845	22	5,262	33
10.1 to 15.0	2,322	13	4,086	26
15.1 to 20.0	524	3	767	5
More than 20	338	2	302	2

[a] Consistent size detail not available for these industries.

The reported change appears to be highly rounded. For other firms, certain observations appear to be outliers—that is, they show enormous, implausible reported or expected price change. While there are only a few outliers, they can significantly affect results especially for disaggregated groups of firms.

What we would like the expectations data to measure is the price expectations that firm managers have in mind when they affect market outcomes; for example, when they conduct wage negotiations or make inventory decisions.[4] It is these price expectations that can reduce firms' profits if they are inaccurate. Furthermore, we would like both the expectations and the actual data to reflect all of the products a firm produces, weighted by relative importance. We are not at all sure that the numbers reported in the BEA survey meet these standards. Survey respondents, typically in the Office of the Controller or in the accounting department, probably know much less about prices than they do about capital expenditures, sales, or other income-statement items. Even when they know about prices, constructing an appropriately weighted average price is a complex task for multiproduct firms. Sizeable departures of data reported by individual firms from the concepts we would like to measure appear likely.

If errors in variables are important, then we would expect bias in regression results and other tests. For example, random errors would cause coefficients to be biased toward zero in regressions of actual on expected price change for individual firms. One remedy for this problem, originally proposed by Abraham Wald (1940), is to group the observations and perform tests on means of grouped data rather than on individual observations. We have therefore constructed data grouped on the basis of year, size of company, industry, and lagged price change, and we have compared a number of regression results for grouped and ungrouped data.

---

[4] For recent evidence consistent with an influence of price expectations on inventory decisions, see de Leeuw.

To construct the grouped data, each observation was classified into one of five years, one of nine industries, one of five size groups, and one of seven categories of lagged actual change in prices (ranging from 0 percent or less to more than 15 percent). All of these variables are predetermined in the relationships we shall be testing, and most of them are probably correlated with "true" price changes, as they should be to be used as a basis for grouping. Except for the lagged actual change, they are all virtually error free —also a desirable characteristic. Lagged actual change is not error free, but the use of ranges instead of exact reported values mitigates the statistical problems of using it as a basis for grouping.

We did not accept grouped observations based on 5 or fewer individual observations. This restriction, we feel, gets rid of virtually all of the outliers in the individual firm data.

The result of our procedure is 606 grouped observations for sales prices and 525 grouped observations for capital goods prices. The median size of groups is 19 individual observations for sales prices and 20 for capital goods prices. If the errors within a group tend to be random, then the errors in grouped observations will tend to vary inversely with the square root of the size of groups. Thus, groups of 16 firms will tend to have errors only one-fourth as large as individual firms.

We compare for grouped and ungrouped data the following regressions:

$$\dot{P}_t = b_0 + b_1\dot{P}_t^e + b_2\dot{P}_{t-1}, \tag{1}$$

where $\dot{P}$ is the actual change in prices in the most recent year, $\dot{P}_t^e$ is the price change expected, as of a year earlier, during year t, and $\dot{P}_{t-1}$ is the actual price change reported a year earlier. If the independent variables are subject to significant errors of measurement, we expect to find different coefficients in the grouped and ungrouped regressions. The particular regressions chosen are ones that test for the bias and efficiency of expectations. If expectations are unbiased and efficient, we expect to find actual and expected price change differing only by a random error term with zero mean, which in turn implies constant terms not significantly different from zero, coefficients not significantly different from 1.0 for the expectation variable and coefficients not significantly different from zero for the lagged actual variable.

The comparison follows (t-statistics are in parentheses):

Sales prices

$$\text{ungrouped: } \underset{(58.2)}{\dot{P}_t = 3.991} + \underset{(40.7)}{.395\dot{P}_t^e} + \underset{(16.0)}{.131\dot{P}_{t-1}}$$
$$\bar{R}^2 = .22,\ n = 17{,}852;$$

$$\text{grouped: } \underset{(14.3)}{\dot{P}_t = 3.158} + \underset{(11.6)}{.630\dot{P}_t^e} + \underset{(0.8)}{.028\dot{P}_{t-1}}$$
$$\bar{R}^2 = .57,\ n = 606;$$

Capital goods prices

$$\text{ungrouped: } \underset{(71.4)}{\dot{P}_t = 5.518} + \underset{(29.5)}{.289\dot{P}_t^e} + \underset{(19.1)}{.181\dot{P}_{t-1}}$$
$$\bar{R}^2 = .23,\ n = 15{,}742;$$

$$\text{grouped: } \underset{(24.8)}{\dot{P}_t = 4.522} + \underset{(12.7)}{.613\dot{P}_t^e} \underset{(-1.2)}{- .043\dot{P}_{t-1}}$$
$$\bar{R}^2 = .71,\ n = 525.$$

The differences between the ungrouped and grouped results are dramatic. For the expected price variable the coefficient rises from .395 (ungrouped) to .630 (grouped) for sales prices, and from .289 (ungrouped) to .613 (grouped) for capital goods prices. For the lagged actual price change, the coefficients for both sales and capital goods prices are positive and highly significant for the ungrouped data, but not significantly different from zero for the grouped data. Because we believe these differences are due to errors in variables, the analysis in the rest of this

TABLE 2—DATA CHARACTERISTICS, BEA GROUPED PRICE DATA

	Sales Prices		Capital Goods Prices	
	Number	Percent	Number	Percent
All Observations	606	100	525	100
Industry:				
Mining	31	5	30	6
Manufacturing	292	48	245	47
Trade	75	12	66	13
Transportation	56	9	45	9
Utilities and Communications	42	7	36	7
Construction and Services	110	18	103	20
Size:				
Manufacturing and Trade:				
Large	98	16	105	20
Medium	184	30	150	29
Small	195	32	159	30
All other	129	21	111	21
Reported Price Change (percent):				
Less than 2	13	2	1	0
2.1 to 4.0	31	5	3	1
4.1 to 6.0	99	16	6	1
6.1 to 8.0	184	30	95	18
8.1 to 10.0	152	25	167	32
10.1 to 15.0	103	17	222	42
More than 15	24	4	31	6

paper will stress grouped data, although we present some ungrouped results for comparison.[5]

The grouped data have distributions differing somewhat from the ungrouped data, as Table 2 shows. By industry, manufacturing now accounts for nearly one-half of the observations instead of one-third; construction and services lose in importance. By size, large firms are more important and small firms less important in the grouped data. By year (not shown in the table), the distribution continues to be even. By reported price change, as we would expect, the variance is much smaller in the grouped data.

III. Tests for Bias in Grouped Data

The regression results just presented suggest that even though the grouped data come much closer than the ungrouped data to passing the test for short-run unbiasedness, they nevertheless still fail to pass the test. In the grouped data as well as in the ungrouped data, constant terms are significantly different from zero. In this section, we examine these relationships for five subsets of the grouped data, dropping the lagged actual price change variable, which did not prove significant in the grouped data. Results appear in Table 3.

For all observations, results are very similar to the ones we have already seen. The table includes an F-test of the joint hypothesis of zero constant term and coefficient of expected change equal to 1.0. The test indicates overwhelming rejection of the hy-

---

[5] Errors in variables is not the only way to account for these differences between the grouped and ungrouped data. Another possibility is that the expected price change $\dot{P}_t^e$ is a much more appropriate dependent variable in these equations than the reported change $\dot{P}_t$, on the grounds that expected price changes may be much more influenced by stochastic factors than actual price changes. Higher coefficients for $\dot{P}_t^e$ in the grouped data could thus reflect less "noise" in the grouped $\dot{P}_t^e$ than in the ungrouped $\dot{P}_t^e$.

If this were the explanation, then rerunning the regressions with $\dot{P}_t^e$ as dependent should *not* give differences between ungrouped and grouped data; with $\dot{P}_t^e$ as dependent, the lower noise level in grouped $\dot{P}_t^e$ should show up simply in the form of smaller regression residuals. When we reran the regressions with $\dot{P}_t^e$ as dependent however, grouped coefficients for $\dot{P}_t$ continued to exceed ungrouped coefficients. Details are available from the authors.

TABLE 3—EXPECTATION BIAS TEST: GROUPED DATA
(Dependent Variable $\dot{P}_t$)

	All Observations	Utilities and Communications	Manufacturing, Mining, Trade			Construction, Transportation, Services
			Large	Medium	Small	
A. Sales Prices						
Constant	3.088	.697	.758	3.138	4.178	2.986
	(15.2)	(0.7)	(1.7)	(7.0)	(15.2)	(7.9)
$\dot{P}_t^e$	.670	1.096	.992	.683	.497	.633
	(28.1)	(11.0)	(16.7)	(12.2)	(16.3)	(13.9)
F-value[a]	190.9^{b}	0.9	0.0	31.9^{b}	271.1^{b}	65.7^{b}
$\bar{R}^2$	.57	.75	.74	.45	.58	.69
Number of Grouped Observations	606	42	98	184	195	87
B. Capital Goods Prices						
Constant	4.603	3.145	3.938	5.196	5.068	4.258
	(27.4)	(4.8)	(9.0)	(15.9)	(16.8)	(10.6)
$\dot{P}_t^e$	.599	.643	.589	.505	.543	.613
	(35.9)	(10.3)	(12.6)	(17.3)	(20.4)	(16.6)
F-value[a]	800.7^{b}	32.4^{b}	77.4^{b}	287.3^{b}	295.1^{b}	109.6^{b}
$\bar{R}^2$	.71	.75	.60	.67	.72	.79
Number of Grouped Observations	525	36	105	150	159	75

Notes: $\dot{P}_t$ refers to the reported price change from year $t-1$ to year t; $\dot{P}_t^e$ refers to the price change expected as of the end of the year $t-1$ from year $t-1$ to year t. The t-ratios are shown in parentheses.

[a] F-values test the joint hypothesis that $a=0$, $b=1$ in the equation $\dot{P}_t = a + b\dot{P}_t^e + u_t$.

[b] F-values that are significantly different from zero at the 95 percent confidence level, indicating rejection of the joint hypothesis.

pothesis of unbiasedness both for all sales prices and for all capital goods prices. For one way of disaggregating the grouped data, however, results are less uniform. The disaggregation separates the data into five categories. The first consists of observations for utilities and communications. The second includes all large firms (firms with over $100 million in assets) in the manufacturing, mining, and trade industries. The third category consists of medium-size firms in manufacturing, mining, and trade (firms with assets between $10 million and $100 million). The fourth consists of small firms in manufacturing, mining, and trade. Finally, the last group consists of all other firms; that is, firms in the construction, transportation, and services industries. This five-way disaggregation will prove illuminating both for the bias tests in this section and for tests of adaptive expectations models to be presented below. Other methods of disaggregation that we tried were much less helpful.[6]

Results of the bias test for the five categories appear in Table 3. For sales prices, the first two groups, utilities and communications firms, and large manufacturing, mining, and trade firms, appear to be unbiased. Sales prices for the other three groups, and capital goods prices for all five groups, in contrast, exhibit systematic bias.

Our conjecture about these differences is that many firms in the first two groups have sufficient market power to have a strong influence over what their sales prices will be in the coming year. In the first group, regulatory agencies also influence prices heavily.

---

[6] Disaggregation by year, for example, revealed only small differences. If anything, expectations appeared to become less "rational" in the later years of the sample period.

The sales price expectations of the utility-communication group and the large manufacturing-mining-trade group are, to a large extent, *plans* about pricing changes to take place in the near future, not *forecasts* of what external market forces are going to bring. The results in the table indicate that the prices which we conjecture represent plans rather than forecasts of market outcomes are unbiased. Actual outcomes differ from expectations only in random fashion.

Sales price expectations of the last three groups of firms, and capital goods price expectations of all firms, represent forecasts of market forces rather than plans. For these expectations we find bias both for sales prices and for capital goods prices.[7] The hypothesis of short-term unbiased expectations is rejected by these data. We therefore turn to the question of whether there is some learning process detectable in these expectations that would serve to limit bias in the long run.

IV. The Formation of Expectations

When we turn from tests for bias and efficiency to an examination of the process by which expectations are formed, the dependent variable becomes expected price change rather than the price change actually reported.

What we are testing for is the presence or absence of processes by which expectational errors would tend to fall to zero in an environment of a steady rate of inflation. The simplest mechanism of this sort is an error-learning model in which the change in the expected rate of inflation depends on the previously observed error. We begin, accordingly, with a test of this hypothesis in the following form:[8]

$$(2) \qquad \dot{P}^e_{t+1} = b_0 + b_1\dot{P}_t + b_2\dot{P}^e_t.$$

The results, using the same groupings as were described in the previous section, appear in Table 4. The error-learning model implies that the constant terms in the equations reported in that table should not be significantly different from zero, and that the sum of the coefficients of the two explanatory variables should not be significantly different from 1.0. As the *t*-test indicates, we reject the latter hypothesis in most cases. For the sales price expectations of regulated firms and large manufacturing, mining, and trade firms, this rejection is not surprising. If it is true, as we conjecture, that sales price expectations for these firms represent plans rather than forecasts of market forces, then it is not surprising that these expectations have little relationship to past expectations and past price changes. Developments and decisions not reflected in the error-learning model surely ought to play a major role in pricing plans. For example, plans for utility firms might be related to the actual or expected outcome of rate decisions, while plans for large manufactures might be related to the pressures of international competition, or to the current state of inventories.

Outside of these two groups of firms, we nevertheless continue to reject the simple error-learning model in most cases. The sum of the coefficients of the two independent variables is less than 1.0 by a significant margin for the three remaining sales price groups, and the constant terms are significantly greater than zero in these groups and in most of the capital goods price groups. An especially puzzling feature of the results is

---

[7] The difference in results between sales prices for large and regulated firms and all other groups helps to explain one feature of the industry aggregates analyzed in our 1981 paper. In that paper, bias tests similar to those presented in Table 3 showed a coefficient significantly less than 1.0 on expectations for capital goods prices, but a coefficient greater than 1.0 for sales prices. Since industry aggregates are very heavily weighted by large firms, the result for sales prices was influenced heavily by firms with market power. The results for capital goods prices were also influenced by large firms, but here there is no difference between large and small firms to make the aggregate different from individual subgroups.

[8] Philip Cagan, in an early and influential statement of the hypothesis (1956), used the form $(\dot{P}^e_{t+1} - \dot{P}^e_t) = c(\dot{P}_t - \dot{P}^e_t)$. This is equivalent to our form if $a = 0$ and $b_1 + b_2 = 1$ (b_1 is then equal to c, and b_2 to $1 - c$). Our form permits testing these two implications of the Cagan form.

TABLE 4—ERROR-LEARNING MODEL: GROUPED DATA
(Dependent Variable $\dot{P}^e_{t+1}$)

	All Observations	Utilities and Communications	Manufacturing, Mining, Trade			Construction, Transportation, Services
			Large	Medium	Small	
A. Sales Prices						
Constant	2.133	3.082	3.845	1.993	1.082	1.276
	(15.9)	(5.1)	(10.7)	(7.6)	(5.5)	(4.5)
$\dot{P}_t$	.569	.561	.392	.488	.824	.748
	(24.9)	(6.0)	(4.8)	(12.7)	(23.8)	(12.2)
$\dot{P}^e_t$	.139	.083	.054	.243	.044	.013
	(6.8)	(.7)	(.6)	(6.2)	(1.9)	(.3)
Sum of $\dot{P}_t$ and $\dot{P}^e_t$	.708	.644	.446	.731	.868	.761
t-ratio[a]	19.0[b]	6.0[b]	13.9[b]	8.6[b]	5.8[b]	7.0[b]
$\bar{R}^2$	.78	.81	.53	.75	.89	.85
Number of Grouped Observations	606	42	98	184	195	87
B. Capital Goods Prices						
Constant	.659	.771	.878	.232	.910	1.200
	(4.9)	(2.0)	(2.6)	(1.0)	(3.3)	(3.2)
$\dot{P}_t$	.961	.972	1.044	1.002	.933	.858
	(42.6)	(12.3)	(18.2)	(27.5)	(21.7)	(12.4)
$\dot{P}^e_t$	−.031	−.051	−.155	−.030	−.017	.014
	(−2.0)	(−.9)	(−3.6)	(−1.3)	(−.6)	(.3)
Sum of $\dot{P}_t$ and $\dot{P}^e_t$	.930	.921	.889	.972	.916	.872
t-ratio[a]	5.5[b]	1.9	3.1[b]	1.3	3.4[b]	3.7[b]
$\bar{R}^2$	.92	.94	.86	.93	.91	.91
Number of Grouped Observations	525	36	105	150	159	75

Notes: See Table 3.
[a] t-ratios test the hypothesis $b_1 + b_2 = 1$ in equation (2).
[b] t-ratios that are significantly different from zero at the 95 percent confidence level, indicating rejection of the hypothesis.

that the coefficient of lagged expectations is negative for all groups of capital goods prices except the last one.[9]

The addition of one more independent variable changes the outcomes shown in Table 4 a great deal. The additional variable is the change in the actual inflation rate from the previous year; that is, $\dot{P}_t - \dot{P}_{t-1}$. The implications of this variable, when it has a positive coefficient, is that expectations are raised not only if the most recent actual price is higher, but also if the most recent actual price has been increasing. Conversely, expectations are slowed down if the recent actual price change has been falling. A significant role for this variable has important implications for the rate at which price expectations adapt to a new underlying rate of inflation. The equation we tested including this new variable is

$$(3)\ \dot{P}^e_{t+1} = b_0 + b_1\dot{P}_t + b_2\dot{P}^e_t + b_3\left(\dot{P}_t - \dot{P}_{t-1}\right).$$

Results based on grouped data appear in Table 5. We continue to reject this augmented error-learning model for the sales prices of regulated firms and large manufacturing, mining, and trade firms. But elsewhere, the results are much closer to a process which would reduce or eliminate expectational bias in the long run. For capital goods prices, this is especially true. The unexplained negative coefficients on expected price change of Table 4 are replaced by positive coefficients in Table 5; and the sum of the coefficients on actual and expected change are in the range of .9 to 1.0. Further-

---

[9] This was a puzzle in our earlier results as well (1981, p. 310).

TABLE 5—AUGMENTED ERROR-LEARNING MODEL: GROUPED DATA
(Dependent Variable $\dot{P}^e_{t+1}$)

	All Observations	Utilities and Communications	Manufacturing, Mining, Trade			Construction, Transportation, Services
			Large	Medium	Small	
A. Sales Prices						
Constant	1.900	3.362	3.811	1.351	.839	.911
	(13.5)	(4.6)	(10.5)	(5.1)	(4.0)	(3.2)
$\dot{P}_t$	.483	.606	.326	.288	.729	.683
	(16.6)	(5.4)	(3.0)	(5.9)	(15.4)	(11.4)
$\dot{P}^e_t$	.264	.003	.132	.552	.175	.138
	(7.9)	(0.0)	(1.0)	(8.7)	(3.4)	(2.5)
$\Delta \dot{P}_t$	.089	−.051	.047	.198	.096	.114
	(4.7)	(−0.7)	(0.9)	(5.9)	(2.9)	(3.7)
Sum of $\dot{P}_t$ and $\dot{P}^e_t$	.747	.609	.458	.840	.904	.821
t-ratio[a]	14.7[b]	5.1[b]	11.0[b]	4.7[b]	3.7[b]	5.0[b]
$\bar{R}^2$	.79	.82	.53	.80	.89	.87
Number of Grouped Observations	606	42	98	184	195	87
B. Capital Goods Prices						
Constant	.409	.239	.510	−.020	.588	.940
	(3.2)	(0.7)	(1.6)	(−0.1)	(2.1)	(2.5)
$\dot{P}_t$	.797	.763	.797	.884	.796	.657
	(28.1)	(9.1)	(10.7)	(19.8)	(14.2)	(6.8)
$\dot{P}^e_t$	.166	.226	.147	.119	.155	.251
	(6.2)	(2.7)	(1.9)	(2.9)	(2.9)	(2.6)
$\Delta \dot{P}_t$	.154	.164	.167	.129	.144	.183
	(8.6)	(4.0)	(4.7)	(4.2)	(3.6)	(2.8)
Sum of $\dot{P}_t$ and $\dot{P}^e_t$	.963	.989	.944	1.003	.951	.908
t-ratio[a]	3.0[b]	.3	1.6	.1	1.9	2.6[b]
$\bar{R}^2$	.93	.96	.88	.94	.92	.92
Number of Grouped Observations	525	36	105	150	159	75

Notes: See Table 3.

[a] t-ratios test the hypothesis $b_1 + b_2$ in equation (3).

[b] t-ratios that are significantly different from zero at the 95 percent confidence level, indicating rejection of the hypothesis.

more, the constant terms for capital goods prices are not significantly different from zero in most groups.

The coefficients imply fairly rapid adaptation of price expectations to a change in the observed inflation rate. The equation for all observations, for example, implies that a rise of 1 percentage point in the observed inflation rate, while it has no contemporaneous effect on the expected inflation rate, causes the expected rate to rise by .95 percentage points in the following year.

For sales prices, the results are a distinct improvement over Table 4, but do not conform as closely to the hypothesis of long-run unbiasedness as the results for capital goods prices. The sum of the coefficients on actual and lagged expected prices are in the range of .8 to .9 rather than the .9 to 1.0 range for capital goods prices, and a t-test of the sums indicates that they remain significantly below 1.0. Since the grouping procedure underlying the regressions does not completely eliminate the effects of errors in variables, the margin below 1.0 could be a remaining effect of the errors. In any case, it seems clear that for both types of prices, as long as we exclude the sales prices of regulated and large manufacturing, mining, and trade firms, the data are consistent with the existence of this augmented error-learning process which would serve to reduce price expectation bias rapidly,

TABLE 6—AUGMENTED ERROR-LEARNING MODEL: UNGROUPED DATA
(Dependent Variable $\dot{P}^e_{t+1}$)

	All Observations	Utilities and Communications	Manufacturing, Mining, Trade			Construction, Transportation, Services
			Large	Medium	Small	
A. Sales Prices						
Constant	2.220	3.939	2.902	2.263	1.760	1.874
	(42.4)	(8.3)	(17.9)	(20.4)	(17.4)	(24.6)
$\dot{P}_t$	.549	.401	.232	.496	.680	.577
	(75.2)	(10.1)	(10.4)	(35.1)	(45.4)	(51.6)
$\dot{P}_t^e$	.165	.194	.366	.200	.112	.181
	(23.3)	(4.5)	(13.9)	(12.6)	(7.9)	(17.7)
$\Delta\dot{P}_t$	.001	−.081	.094	.035	.043	−.046
	(0.3)	(−2.7)	(5.5)	(2.9)	(3.5)	(−5.2)
Sum of $\dot{P}_t$ and $\dot{P}_t^e$	.714	.595	.598	.696	.792	.758
t-ratio[a]	49.7[b]	11.6[b]	20.4[b]	22.4[b]	19.0[b]	28.5[b]
$\bar{R}^2$	.51	.32	.45	.51	.63	.53
Number of Observations	17852	693	1377	3679	4134	7969
B. Capital Goods Prices						
Constant	1.488	2.180	1.452	1.162	1.618	1.655
	(24.6)	(9.5)	(9.0)	(10.1)	(11.1)	(16.6)
$\dot{P}_t$	.694	.551	.707	.741	.710	.671
	(89.2)	(16.6)	(30.5)	(48.0)	(40.7)	(53.5)
$\dot{P}_t^e$	.167	.222	.136	.141	.143	.184
	(24.4)	(7.3)	(6.4)	(10.4)	(8.9)	(17.2)
$\Delta\dot{P}_t$	.066	.097	.073	.067	.047	.068
	(10.2)	(4.5)	(4.3)	(5.1)	(3.1)	(6.6)
Sum of $\dot{P}_t$ and $\dot{P}_t^e$	.861	.773	.843	.882	.853	.855
t-ratio[a]	24.7[b]	9.7[b]	9.3[b]	10.8[b]	11.4[b]	16.6[b]
$\bar{R}^2$	.65	.59	.67	.70	.65	.63
Number of Observations	15742	838	1582	3614	3062	6646

Notes: See Table 3.

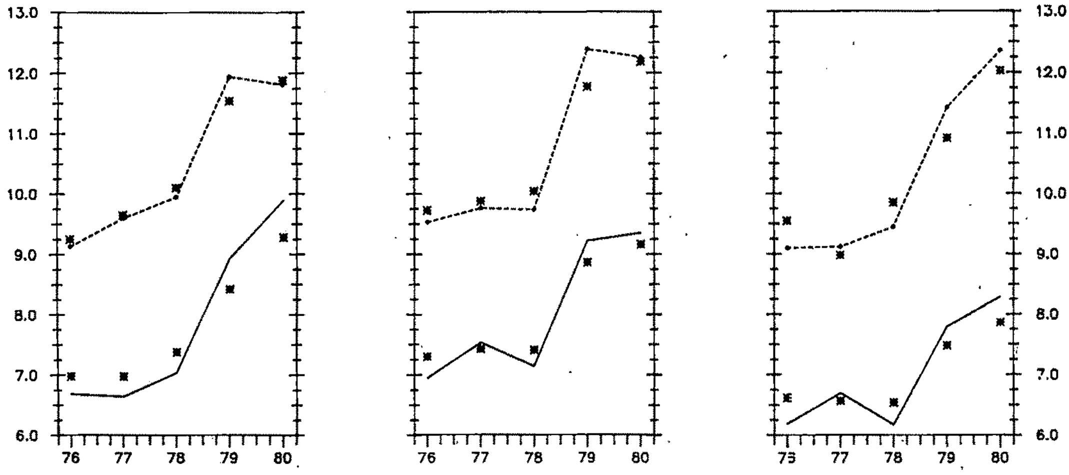
[a] t-ratios test the hypothesis $b_1 + b_2$ in equation (3).

[b] t-ratios that are significantly different from zero at the 95 percent confidence level, indicating rejection of the hypothesis.

if not quite completely, in the presence of a steady rate of inflation.

Long-run bias persists, however, in results based on individual firm data rather than grouped data. Individual firm results for the augmented error-learning model appear in Table 6. For sales prices of regulated and large firms, the coefficients of $\dot{P}_t^e$ and $\dot{P}_t$ sum to far below 1.0 for both grouped data (Table 5) and ungrouped data (Table 6). For other categories, the sum of these coefficients is systematically lower in Table 6 than in Table 5. For example, for medium-size firms in manufacturing, mining, and trade, the sum of coefficients based on grouped data is .840 for sales prices and 1.003 for capital goods prices; the corresponding sums based on ungrouped data are .696 and .882. Rejection of the long-run bias proposition evidently depends on choosing the grouped over the ungrouped results.

When we examine the residuals from regressions based on grouped data (Table 5) to see if they contain any systematic features, we find that residuals grouped by year are not random. Year-by-year average actual and predicted values are plotted in Figure 1 for three categories: medium-size manufacturing-mining-trade; small manufacturing-mining-trade, and construction-transportation-services. The year-by-year average residuals are not large, but they display a similar pattern in the six cases. They are all negative in 1976, mixed in 1977, all negative



(a) Medium-size firms, manufacturing-mining-trade

(b) Small firms, manufacturing-mining-trading

(c) Construction-transportation-services

FIGURE 1. EXPECTED PRICE CHANGES REPORTED BY FIRMS AND PREDICTED BY AUGMENTED ERROR-LEARNING MODEL: AVERAGES BY YEAR, GROUPED DATA

Note: _____Expected sales price changes reported by firms

_ _ _ _Expected capital goods price changes reported by firms

* * * *Expected price changes predicted by augmented error-learning model

in 1978, all positive and relatively large in 1979, and all but one positive in 1980. Evidently some force or forces not covered by the augmented error-learning process were lowering expectations in 1976 and 1978, and raising them in 1979 and 1980.

Obvious candidates for the missing variables are indicators of policy changes, or of pressures of demand against normal supply.[10] We tested three variables that are often discussed in connection with inflation; the money supply, the federal deficit, and materials capacity utilization. For money, the specific variable was the percent change in *M*1 (formerly *M*1*B*) for the year ending in the third quarter just before price expectations were reported. For the deficit, we tested the ratio of the latest fiscal year federal deficit to *GNP* in four forms: the actual ratio, the ratio using the high-employment deficit, and the changes in these two ratios. For materials capacity utilization, we tested the seasonally adjusted rate for the third quarter just before price expectations were reported. We expected a positive effect on price expectations for all of these variables.

None of the variables had the expected effect. The money supply and utilization variables were not significant, while the deficit variable in most forms was significant with a negative, rather than a positive, sign. In lagged form, however, two of the variables—the change in the money supply and the level of capacity utilization—had expected positive signs. Of course, with only five observations for these aggregate variables (one for all 1976 observations, one for all 1977 observations, etc.), it is hardly surprising that some transformation of some one of the variables will have the expected sign. But perhaps it can be concluded that the BEA price expectations data are at least not inconsistent with some direct influence of lagged money supply changes and lagged capacity utilization on price expectations. The size of the coefficients suggests that such direct influence, if it exists, is small.[11]

---

[10] Gramlich tests for the influence on household price expectations of indicators of policy, demand pressure, and other variables. A number of the variables he considers, including indicators of "policy credibility" cannot be tested for the short timespan that our data cover.

V. Conclusions

The principal goal of this paper was to see if there are any signs in the BEA data on price expectations of a process of expectation formation which would serve to reduce bias in the long run, so that a 1 percent increase in actual price change would eventually lead to a 1 percent increase in expected price change. We conclude that, after two major adjustments in the data and one addition to a simple error-learning model, there are signs of such a process.

The adjustments in the data are (a) substituting grouped observations for individual firm observations and (b) omitting sales price observations for communications and utility firms and for large manufacturing, mining, and trade firms. The first adjustment is justified, we believe, because of the strong possibility of major errors in individual firm data. We defend the second adjustment on the basis that sales price expectations for the excluded regulated and large firms may well not represent forecasts of the outcome of market forces, but rather company plans for determination of own prices.

After these adjustments, the data seem consistent with a process in which expected price change is related to the previous expected price change, the most recent observed price change, and the change in the most recent observed price change. The sum of coefficients of the first two variables should equal 1.0 to indicate unbiased expectations in the long run. The actual sums are fairly close to 1.0; they cluster around .95 for capital goods prices and around .85 for sales prices. In response to a step change in the actual inflation rate, the coefficients imply that firms typically make substantial short-run expectational errors, but that these errors are largely eliminated after a year or two.

---

[11] Detailed regression results are available from the authors on request.

REFERENCES

Cagan, Philip, "The Monetary Dynamics of Hyperinflation," in M. Friedman, ed., *Studies in the Quantity Theory of Money*, Chicago: University of Chicago Press, 1956.

de Leeuw, Frank, "Inventories and Economic Instability," *Survey of Current Business*, December 1982, *62*, 25–31.

______ **and McKelvey, Michael J.,** "Price Expectations of Business Firms," *Brookings Papers on Economic Activity*, 1: 1981, 299–314.

Gramlich, Edward, "Models of Inflation Expectations Formation: A Comparison of Household and Economist Forecasts," *Journal of Money, Credit and Banking*, May 1983, *15*, 155–73.

Juster, F. Thomas and Comment, Robert, "A Note on the Measurement of Price Expectations," IRS Working Paper, University of Michigan, 1980.

Leonard, Jonathan S., "Wage Expectations in the Labor Market: Survey Evidence on Rationality," Working Paper 440, National Bureau of Economic Research, 1980.

McKelvey, Michael J. and de Leeuw, Frank, "Price Reports by Business Firms: Comparison with Official Price Indexes," *Review of Public Data Use*, 1982, *10*, 141–51.

Wald, Abraham, "The Fitting of Straight Lines if Both Variables are Subject to Error," *Annuals of Mathematical Statistics*, 1940, *11*, 284–300.

U.S. Office of Business Economics, Bureau of Economic Analysis, "Survey of Plant and Equipment Expenditures," *Survey of Current Business*, various issues.

Models of Arbitrator Behavior: Theory and Evidence

By ORLEY ASHENFELTER AND DAVID E. BLOOM*

The arbitration of disputes by third-party neutrals has become a major industry in the United States. Arbitration's use in the settlement of disputes arising under existing contracts ranges from the settlement of grievances in union-management contracts to the settlement of highly complex disputes between buyers and sellers in commercial contracts. Arbitration is also now widely used to arrive at the terms of new contracts. The use of these provisions for contract, or "interest," arbitration ranges from the highly public disputes between baseball players and club owners to the settlement of public sector labor disputes and the fashioning of divorce settlements.

Since arbitration mechanisms may be designed in several ways, their increased use makes it increasingly important to understand how the differences in arbitration systems may affect outcomes. Our purpose in this paper is to open up the empirical analysis of some simple models of arbitrator behavior under alternative mechanisms and in different economic environments. We do this by studying the outcomes in the first three years of operation of a New Jersey statute that mandates the arbitration of unsettled pay disputes by New Jersey police officers and the municipalities that employ them. This remarkable statute provides for conventional arbitration of pay disputes if the two parties can agree to this, but requires the use of final-offer arbitration if they cannot. Consequently, the results of both mechanisms may be analyzed and compared.

In proceeding this way we hope to shed light on two related issues. First, arbitration mechanisms are simply one specialized type of legal setting for the settlement of disputes. The setting is analogous to a civil suit with the arbitrator acting as judge. Indeed, each side's "case" is often presented to the arbitrator by the parties' attorneys, and the arbitrator produces a written decision after hearing the case. Negotiated settlements are entirely analogous to "out of court" settlements in civil suits, and arbitrated cases are the ones that "go to trial." What is unusual about arbitration systems is that the arbitrator's decisions, and sometimes the proposals made by the parties, are both publicly obtainable and easily reduced to a single numerical magnitude. Consequently, we suspect that a great deal may be learned about the general nature of legal decision making by quantitative studies of the simpler arbitration prototypes.[1]

Second, the growing theoretical literature on the nature of alternative arbitration mechanisms has raised a number of issues whose resolution requires empirical inquiry. In conventional arbitration, for example, the arbitrator fashions an award based on an analysis of the relevant facts and the arbitrator's external judgment of what would comprise a fair award. It has often been suggested that this arbitration system generates little useful information from the parties, and causes the parties' settlement offers to diverge, because to do otherwise might prejudice the arbitrator's judgment against the parties' self interests. To remedy this potential difficulty, Carl Stevens suggested a system of final-offer arbitration nearly two decades ago (1966). He suggested that in order to induce settlement, each party should be required to submit to the arbitrator a single final offer, and the arbitrator would then be required to select one or the other of those offers *without compromise*.[2]

---

*Professor, Department of Economics, Princeton University, Princeton, NJ 08544, and Assistant Professor, Department of Economics, Harvard University, Cambridge, MA 02138, respectively. We are grateful for useful discussions with Christopher Cavanagh, Henry Farber, Richard Lester, Daniel McFadden, and Jeffrey Tener, and comments by Vincent Crawford and John Pencavel on an early draft. Daniel Rosenblum provided expert computational assistance. Bloom's research was supported, in part, by a grant from the Ford Motor Company to Carnegie-Mellon University, where Bloom was on the faculty of the School of Urban and Public Affairs.

[1] For a similar view, see especially William Landes and Richard Posner's wide ranging discussion of private adjudication systems (1979), and the paper by John Gould (1973).

In early theoretical work, Vincent Crawford (1979) observed that if the arbitrator's (exogenously determined) notion of a fair settlement were known to the two parties, then both arbitration mechanisms would almost certainly lead to the same outcome in a zero-sum setting. In conventional arbitration, this would happen because whatever negotiated settlement was proposed by one party, the other party would always do better by holding out for the arbitrator's "fair" decision. Thus, all outcomes would coincide with the arbitrator's preference either through negotiation under the threat of arbitration or by arbitrator decision. Assuming the arbitrator selects the final offer closest to his notion of a fair settlement, this would also be the case with final-offer arbitration. Again, whatever proposal was made by one party, the other party could always do better by holding out for the arbitrator's preferred outcome and both parties would inevitably be led to agree on this outcome or see it imposed by the arbitrator. In this scenario, the threat of arbitration effectively determines the outcome of all settlements and, if the parties bear the arbitration costs, the actual arbitration of cases would wither away.[3] Although this might appear to the outsider as the great success of the arbitration system in encouraging the parties to bargain on their own, precisely the contrary would be the case. The parties would merely be agreeing to the outcome that is already a foregone conclusion and saving themselves the cost of arbitration.

The key to Crawford's conclusion is, of course, the assumption that both parties know with certainty the arbitrator's preferred outcome. In a series of papers, Henry Farber (1979, 1980) and Farber and Harry Katz (1979) have explored the case where the parties are uncertain as to the arbitrator's preferences. In this setting, the outcomes under conventional and final-offer arbitration will generally differ. Negotiated settlements (and whether these will occur) will then depend, in different ways under the two arbitration systems, on the parties' attitudes toward risk, their prior views on the distribution of arbitrator preferences, and other variables.

Distinguishing whether it is more or less reasonable to assume that the parties are certain about arbitrator preferences is equivalent to testing whether the arbitrator's decisions under a final-offer statute are predictable once information on the final offers and the environment of the bargaining unit is known. In this paper we put these issues to an empirical test in an effort to discriminate between the implications of these two theoretical setups. In order to do this, we set out simple models of arbitrator behavior under both final-offer and conventional arbitration, and provide a method for testing the empirical implications of these models for observed arbitration awards and for the "fairness" of arbitrator behavior.

I. Models of Arbitrator Behavior

The New Jersey Fire and Police Arbitration Act was approved in May 1977. The establishment of some kind of arbitration statute followed almost inevitably from the passage in 1968 of legislation that granted New Jersey's public sector employees the right to organize and bargain collectively, but not the right to strike. The difficulty was that in the absence of the right to strike, negotiations often went on long after annual municipal budget cycles had been closed. The purpose of the arbitration statute was to ensure that contract negotiations were final by the time of the employer's budget submission date.

---

[2] Stevens' remarkable paper raises most of the important issues relevant to the analysis of this arbitration scheme. The use of his system is growing and has already been adopted for the settlement of certain public sector labor disputes in six U.S. states and in Chile, and has also been used in settling disputes in major league baseball. James Meade, a prominent member of Britain's new Social Democratic Party, has suggested its use in Great Britain as the enforcement mechanism for a new wage and price monitoring system.

[3] See Bloom (1981) for a more complete analysis of the role of arbitration costs.

To accomplish this purpose, employees and employers covered by the Act (mainly police and fire department workers) must begin collective bargaining at least 120 days before the employer's budget submission date. If the parties do not reach an agreement by 60 days prior to this date, they may select the conventional arbitration mechanism for settlement of the unresolved issues. If the parties cannot or prefer not to agree to conventional arbitration, they are then compelled to have the disputed issues resolved by final-offer arbitration, with the economic issues taken as a single package.[4]

Arbitrators are assigned to cases by the New Jersey Public Employment Relations Commission (PERC), which is responsible for administration of the Act. A list of seven members from the PERC's special panel of arbitrators is first circulated to the parties for comment. One of the seven is then appointed to the case, with all arbitration fees borne jointly by the parties (subject to a fee schedule approved by the PERC).

A. *Arbitrator Behavior under Final-Offer Arbitration*

Under final-offer arbitration, the arbitrator is required to select either the union's final offer w^u or the employer's final offer w^e as the settlement.[5] In order to do this, we suppose that the arbitrator has in mind a preferred settlement w^a. The value of w^a is presumably based on the application of subjective principles of equity to the objective considerations of a particular case.[6] The value of w^a in any particular arbitration case is, of course, unknown to outside observers. Moreover, in a case resolved by final-offer arbitration, w^a will never be revealed.

The preferred settlements of arbitrators w^a will presumably vary across arbitration cases both because of unobserved variations in economic environments and because of differences in arbitrators' assessments of those circumstances. We shall suppose that in a given year, w^a may be thought of as being drawn from an as yet unspecified distribution function. The simplest mechanism by which an arbitrator in a particular case might select an offer is to choose that party's offer that is closest to w^a. In these circumstances, the arbitrator would select the employer's offer if

$$(1) \qquad |w^a - w^e| \leq |w^u - w^a|,$$

and select the union's offer otherwise. Since $w^e < w^u$ or there would be no need for arbitration, (1) implies that the employer's offer is selected if

$$(2) \qquad w^a \leq (w^e + w^u)/2.$$

Given the established values of the employer and union final offers, (2) implies that the employer's offer will be chosen if an arbitrator is selected for the case whose preferred settlement is less than the simple average of w^e and w^u. It follows that the probability of an employer victory, P, is simply

$$(3) \qquad P = \text{Prob}[w^a \leq (w^e + w^u)/2] = F[(w^e + w^u)/2],$$

where $F(\cdot)$ indicates the value of the cumulative distribution function that describes arbitrator preferences.[7] If we take the distribution of arbitrator preferences to be normal with mean μ^p and standard deviation σ^p, as we shall do in our empirical work, (3) is simply

---

[4] We analyze only disputes in police work in this paper, as many of New Jersey's fire departments are volunteer companies. Other details of the statute are discussed by Bloom (1980).

[5] Under the New Jersey statute, those issues that may be reduced to a single numerical award are the "economic" issues, and it is these that we examine throughout. "Noneconomic" issues are handled as a separate item under the New Jersey statute, and their analysis raises some difficult issues. In fact, however, most disputes involving police officers in New Jersey have been over pay, so that our focus on economic issues is probably the appropriate one.

[6] In New Jersey, arbitrators are directed to give due weight to an enormously broad variety of factors that includes 1) the interests of the public, 2) relevant comparisons of wages and working conditions, 3) the overall level of compensation already received, 4) the financial impact of the settlement on the municipality, 5) the cost of living, and 6) the continuity of employment. Considerable discretion is obviously left to individual arbitrators.

[7] Although his interpretation is different, equation (3) is identical to Farber's (1980) formulation of a similar problem. Our procedure was to some extent inspired by his.

$$(4)\quad P = N[(w^e + w^u)/2\sigma^p - (\mu^p/\sigma^p)],$$

where $N(\cdot)$ indicates the value of the cumulative standard normal distribution function. Expression (4) is nothing more than a simple probit function, and its parameters are straightforward to estimate by standard maximum-likelihood methods. The constant in this probit function is an estimate of μ^p/σ^p, while the coefficient of $(w^e + w^u)/2$ is an estimate of $1/\sigma^p$. It follows that although the various realized values of w^a may not be observed, the parameters of the distribution function from which the arbitrator's preferences are drawn may be estimated from a series of observations on union and employer final offers, and arbitrator choices from those offers.

It is worth observing that the evenhanded arbitrator treatment of union and employer offers embedded in the decision rule (1) has some immediate superficial plausibility, at least under the New Jersey statute. After all, arbitrator fees are borne by the parties *and* the parties' choices are taken into account in the selection of an arbitrator from the panel of seven arbitrators initially proposed. Since arbitrators derive considerable income from these fees, the decision rule (1) may be the best viable strategy for maximizing an arbitrator's income over any long-run period.[8]

A major objection to the evenhanded arbitrator treatment implied by (1) is that it rules out arbitrator bias as might be indicated by the asymmetric treatment of union and employer final offers. As it turns out, the presence of this form of arbitrator bias may be tested for empirically. Suppose, for example, that the employer's offer is accepted if

$$(5)\quad \alpha(w^a - w^e) < (w^u - w^a),$$

where $\alpha > 0$. In this setup, arbitrators demonstrate a pro-union, pro-employer, or no bias as $\alpha > 1$, $\alpha < 1$, or $\alpha = 1$, respectively. With this specification, the probability that an employer's offer is accepted is

$$(6)\quad P = N\{[\alpha/(1+\alpha)\sigma^p]w^e + [1/(1+\alpha)\sigma^p]w^u - \mu^p/\sigma^p\}.$$

Equation (6) does nothing more than free up the coefficients on w^e and w^u in the probit function. The ratio of these coefficients is an estimate of α, while their sum is an estimate of $1/\sigma^p$. We shall test the hypothesis $\alpha = 1$ in the empirical work below.

An alternative objection to the decision rule (1) is that it implies that arbitrator notions of fairness do not explicitly take account of the parties' final offers. It is not difficult to see, however, that so long as the arbitrator weighs the employer and union offers symmetrically, the decision rule (1) is observationally equivalent to an entire class of rules where the arbitrator does take these offers explicitly into account. Suppose, for example, that the arbitrator's preference $\tilde{w}^a$ is formed as

$$(7)\quad \tilde{w}^a = \gamma w^a + (1-\gamma)(w^e + w^u)/2,$$

with $0 < \gamma < 1$. Here $\tilde{w}^a$ is a weighted average of w^a, the arbitrator's preference in the absence of information on the offers, and the mid-point of the final offers. The employer's offer is accepted if $|\tilde{w}^a - w^e| < |w^u - \tilde{w}^a|$ which will be the case if

$$\tilde{w}^a = \gamma w^a + (1-\gamma)(w^e + w^u)/2 < (w^e + w^u)/2,$$

or if

$$(8)\quad \gamma w^a < \gamma(w^e + w^u)/2.$$

Condition (8) is, of course, identical to (1). From an empirical point of view, therefore, there is an infinite set of symmetric arbitrator decision-making processes that all lead to the same observationally equivalent decisions. In this sense, the decision rule (1) may be a very robust empirical device.

---

[8] This view of the arbitrator is also suggested by Otto Davis (1979). As this discussion indicates, however, the incentive structure for arbitrators is an issue that deserves further research. David Card (1983) has begun it in his study of grievance arbitration.

It is worth observing that the structure of the empirical analysis varies slightly when arbitrators use decision rule (5) and when arbitrator notions of fairness are formed according to equation (7). In this case, the probability that an employer's offer is accepted is

$$P = N\{\delta_1 w^e + \delta_2 w^u - \mu^p/\sigma^p\}, \tag{9}$$

where $\delta_1 = [\alpha/(1+\alpha) - (1-\gamma)/2]/\gamma\sigma^p$

and $\delta_2 = [1/(1+\alpha) - (1-\gamma)/2]/\gamma\sigma^p$.

By comparison, it is clear that equations (9) and (6) are observationally equivalent and that α and γ are not separately identified in the latter formulation. Nevertheless, as with (6), the sum of the coefficients of w^e and w^u in (9) is an estimate of $1/\sigma^p$, implying that μ^p is identified, and the equality of the coefficients δ_1 and δ_2 continues to provide a test of the hypothesis $\alpha = 1$.

The test of the hypothesis $\alpha = 1$ is mainly a test of the decision rule (1) under the maintained hypothesis that the distribution function characterizing the w^a may be taken to be normal. To this point, however, we have not attempted to characterize this distribution in any more detail. For many observers, however, a test of the fairness of arbitrators is really a test of whether μ^p is "fair," or of whether arbitrators have the "correct" preferences. Of course, it is difficult to specify the value for μ^p that would be fair, but it is not difficult to make μ^p a function of some specified set of variables whose coefficients can be estimated. For the ith observation, we may write $\mu_i^p = x_i\beta^p$ to get the probit function

$$P_i = N\{[\alpha/(1+\alpha)\sigma^p]\, w_i^e + [1/(1+\alpha)\sigma^p]\, w_i^u - x_i\beta^p/\sigma^p\}. \tag{6'}$$

This merely introduces the variables x_i into the probit function. Our estimates of the coefficients in the vector β^p will indicate how the variables in the vector x determine the mean of the distribution of the arbitrators' notions of what is fair. These estimates may, of course, be compared by the outside observer against any proposed definition of what should determine on average a fair award. Indeed, the econometric analysis that follows may be interpreted as nothing more than a scheme for uncovering the nature of arbitrator preferences when only qualitative information on arbitrator awards is available.

We consider three main candidate variables to enter the vector x_i. First, it seems likely that a measure of the wage rate in some alternative occupation may influence an arbitrator's decision. If the relative supply of workers to police departments were very elastic with respect to this alternative wage, the wage structure determined in a competitive market would, of course, guarantee one-to-one movements in the wage rates of police workers and the wage rates in their alternative occupations. For a variety of reasons, we doubt that the relative supply of police workers is highly elastic in the short run, but the hypothesis that arbitrators form their notions of a fair award as if it were true has considerable intuitive appeal and seems worth testing. Second, it also seems possible that arbitrators may act to narrow (or occasionally widen) the wage differentials of police workers across municipal jurisdictions. To the extent that this occurs, we may expect that an arbitrator's notion of a fair wage increase will be negatively (or occasionally positively) correlated with the level of the wage in the previous contract in that municipality. To test this hypothesis we include the level of the previous contract wage as an element of the vector x_i. Finally, we consider the impact of the municipality's financial well-being on the arbitrator's decision. We include in the vector x_i both the municipality's tax rate and its per capita debt, with the presumption that greater values of both of these variables operate to reduce the arbitrator's notion of a fair award. We conclude by observing that, as we have already noted, the New Jersey statute directs arbitrators to consider all of these items in framing their decisions, but without stipulating how or in what form. It is an empirical question, therefore, as to whether any or all of these variables bear a stable relationship to arbitrator preferences.

B. *Arbitrator Behavior under Conventional Arbitration*

Under conventional arbitration, the arbitrator is free to fashion a settlement of his own choosing. As before, we assume that w^a represents the arbitrator's preferred award in the absence of information on the parties' proposed offers. We consider two alternative schemes by which the arbitrator may fashion and then impose a settlement.[9]

The simplest scheme would be for the arbitrator to simply impose his preferred award w^a as the settlement. In this case, observations on conventional arbitration awards would directly reveal arbitrator preferences. For the ith arbitration case, we would then have

$$w_i^a = \mu^r + \varepsilon_i, \tag{10}$$

where ε_i has mean zero and standard deviation σ^r. If the mean of the distribution of the arbitrator's preferred settlements varies with some vector of variables x_i whose coefficient vector is β^r, this would lead to nothing more than the regression function

$$w^a = x_i\beta^r + \varepsilon_i, \tag{11}$$

which is easily fit to the data on conventional arbitration awards.

Alternatively, it may be supposed that the arbitrator attempts to fashion a compromise based on w^a and the arguments presented to him by the parties. Suppose that the last offers presented by the parties to each other are w^e and w^u, and that these are made known to the arbitrator. A natural compromise arbitration award would then be

$$\begin{aligned} \tilde{w}_i^a &= \rho w_i^a + (1-\rho)(w_i^e + w_i^u)/2 \\ &= \rho(x_i\beta^r) + (1-\rho)(w_i^e + w_i^u)/2 + \rho\varepsilon_i \\ &= \rho(x_i\beta^r) + v_i, \end{aligned} \tag{12}$$

where $v_i = (1-\rho)(w_i^e + w_i^u)/2 + \rho\varepsilon_i$,

and $0 < \rho < 1$.

The second line of (12) might also be fit directly as a regression function to the data if w^e and w^u were observed. In conventional arbitration proceedings, however, these offers are not generally known to outside observers, and they may never be stated explicitly by either party in any case. If the unobserved offers w^e and w^u are correlated with the variables in the vector x_i, as may well be the case, our regression estimates of $\rho\beta^r$ in (12) will suffer from conventional omitted variable bias.

The question naturally arises as to how the models of conventional arbitrator behavior in (11) and (12) might be distinguished empirically. It should be clear that this will not generally be possible without observations on w^e and w^u when all arbitrators are assigned to conventional arbitration cases. In New Jersey, however, the same panel of arbitrators is used in both conventional arbitration and final-offer arbitration cases. It follows that for this group of arbitrators, we may assume $\mu^r = \mu^p$, $\sigma^r = \sigma^p$, and $\beta^r = \beta^p$. Under these assumptions, we may fit equations (6′) and (11) separately and use a test for the equality of these parameters as evidence to favor one or the other of the formulations of (11) or (12). If, for example, the inverse of the standard deviation of conventionally arbitrated awards is equal to the slope of the probit function in (4), we have evidence that arbitrators simply impose upon the parties their notion of a fair award, as in (10). If, on the other hand, this equality is not satisfied in the data we may have evidence that the arbitrators take some account of the unobserved offers of the parties, as in (12).[10]

II. Empirical Results

The likelihood function for the final-offer arbitration sample is simply

$$L^p = \prod_{i \in e} P_i \prod_{i \in u} (1 - P_i), \tag{13}$$

---

[9] Both of these are possibilities suggested by Farber (1981).

[10] There is one disturbing case where this test may give misleading results. Suppose that (12) is the correct structure for arbitrator preferences, but that the estimates of (11) are biased by the omission of the variable $(w^e + w^u)/2$ so as to cause acceptance of the cross-equation constraints. Although this seems unlikely, the positive correlation that presumably exists between $x_i\beta^r$ and $(w^e + w^u)/2$ will tend to counteract the attenuating effect of ρ on the parameters estimated (since $0 < \rho < 1$).

TABLE 1—MEANS AND STANDARD DEVIATIONS OF FINAL-OFFER AND CONVENTIONAL ARBITRATION AWARDS, POLICE ARBITRATION CASES, NEW JERSEY, 1978–80[a]

	1980	1979	1978
Mean of Employer Offers ($100\cdot w^e$)	5.70 (2.57)	6.51 (1.35)	5.01 (1.44)
Mean of Union Offers ($100\cdot w^u$)	8.54 (1.46)	8.29 (1.73)	7.14 (1.12)
Mean of Final-Offer Arbitration Awards	8.10 (1.41)	7.57 (1.48)	6.63 (1.19)
Number of Final-Offer Arbitration Cases	109	92	123
Proportion of Employer Victories	.266	.348	.317
Mean of Conventional Arbitration Awards	8.26 (2.14)	8.59 (2.32)	6.55 (2.23)
Number of Conventional Arbitration Cases	32	20	47
Proportion of Bargaining Cases Going to Arbitration	.30	.34	.49

Source: Authors' tabulation of arbitrator reports.
[a] Standard deviations are shown in parentheses.

where P_i is given by (6) (or (6′)) and the first product $\Pi_{i\in e}P_i$ is taken over employer victories and the second product $\Pi_{i\in u}(1-P_i)$ is taken over union victories. The likelihood function for the conventional arbitration sample is simply the likelihood for a conventional regression,

$$(14)\qquad L^r=\Pi(1/\sigma^r)n\left[(w_i^a-x_i\beta^r)/\sigma^r\right],$$

where $n(\cdot)$ indicates the standard normal density function. So long as the parameters β^j and $\sigma^j(j=p,r)$ are taken to be different, (13) and (14) may be maximized independently by the usual procedures. In the case where $\mu=\mu^r=\mu^p$, or $\beta=\beta^r=\beta^p$, and $\sigma=\sigma^r=\sigma^p$, however, the likelihood of the combined sample is

$$(15)\qquad L=L^p\cdot L^r,$$

and the parameters μ, β, and σ are common to all parts of the likelihood function.[11]

A. *Basic Results*

Sample statistics for the basic data are given in Table 1. These data were collected directly from written arbitration reports. In a few cases, all the necessary data were not available in a given report and the observation had to be deleted. Arbitration awards and final offers are given throughout in the form of percentage increases in total compensation.

The sample statistics in Table 1 reveal much of what the econometric analysis will confirm. First, the mean of the union and employer final offers is quite low when compared to the mean of the conventionally arbitrated settlements. In 1979 and 1980, the mean of the union offers is about equal to the mean of the conventionally arbitrated settlements, while the mean of the employer offers is more than two percentage points lower. Although not quite so extreme in the data for 1978, this phenomenon shows up there also. If the mean of the conventionally arbitrated cases is taken to indicate the mean of the distribution of arbitrators' preferred settlements, this suggests that the union groups were behaving very conservatively indeed. Second, consistent with this presumption is the fact that in all three years, the union offers were accepted in roughly two-thirds to three-quarters of the cases. The result was that actual awards under conventional and final-offer arbitration were not nearly so far apart as the parties' final offers might indicate. Finally, the proportion of

---

[11] The likelihood function (15) is identical to that used by Ashenfelter (1983) for an altogether different purpose, and its maximization was carried out by similar methods.

TABLE 2—ESTIMATES OF EQUATIONS (6) AND (10), THE DETERMINANTS OF THE PROBABILITY OF EMPLOYER VICTORIES, POLICE ARBITRATION CASES, 1978–80[a]

	1980				1979				1978			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
α	1.0[b]	1.0[b]	2.010	1.479	1.0[b]	1.0[b]	0.832	0.745	1.0[b]	1.0[b]	−22.588	−3.902
			(.835)	(.356)			(.109)	(.194)			(6.083)	(.627)
μ^p	8.268	8.431	8.259	8.142	8.505	8.476	8.536	8.677	7.407	6.937	6.361	5.709
	(.255)	(.224)	(.244)	(.294)	(.421)	(.303)	(.285)	(.505)	(.381)	(.219)	(3.059)	(3.126)
σ^p	1.481	1.942	2.006	1.567	2.536	2.349	2.352	2.624	2.701	2.234	2.244	2.258
	(.224)	(.189)	(.205)	(.273)	(.456)	(.279)	(.259)	(.595)	(.494)	(.244)	(1.221)	(.418)
μ^r	8.255	8.431	8.259	8.255	8.487	8.476	8.536	8.587	6.550	6.937	6.361	6.550
	(.371)			(.371)	(.506)			(.501)	(.322)			(.322)
σ^r	2.101	1.942	2.006	2.101	2.265	2.349	2.352	2.265	2.211	2.234	2.244	2.211
	(.262)			(.263)	(.358)			(.352)	(.228)			(.234)
−Log Likelihood	116.62	117.65	116.85	116.36	90.79	90.89	90.86	90.76	174.96	176.34	171.88	171.14
Number of FOA Observations	109	109	109	109	82	82	82	82	120	120	120	120
Number of CA Observations	32	32	32	32	20	20	20	20	47	47	47	47

Note: FOA indicates final-offer arbitration cases; CA indicates conventional arbitration cases.

[a] Estimated standard errors are shown in parentheses.

[b] $\alpha = 1.0$ is imposed in the estimation.

cases reaching arbitration dropped considerably between 1978 and 1980. This is consistent with the notion that the use of arbitration will decline considerably as more is learned by the parties about arbitrator preferences. Between the first and second years of the statute's operation in New Jersey, the percentage of negotiations that ended up in arbitration declined from nearly 49 to 34 percent. The drop in the utilization of arbitration from the second to the third year of the statute's life was far less dramatic, and it is our casual impression that further declines have been smaller still. This suggests, although the data are not yet available to confirm it, that at least one-quarter of negotiations may continue to end in arbitration even after the parties are thoroughly familiar with the statute's operation.

Table 2 contains the empirical results obtained by maximizing various versions of the likelihood functions (13)–(15) when the variables x_i are ignored. These results are perhaps the easiest to interpret and they display much of what the data contain. In column (1) we present the simplest results for each year that correspond to the probit function (4) and the regression equation (10). In these results, we impose the constraint $\alpha = 1.0$. Consider first the results for 1980. The mean of arbitrator preferences from the final-offer arbitration cases (μ^p) is estimated at 8.3 percent, which is considerably higher than the mean of the final offers of the employer groups and only slightly lower than the mean of the final offers of the union groups indicated in Table 1. The standard deviation of arbitrator preferences from the final-offer arbitration cases (σ^p) is estimated at 1.5 percentage points and, as (4) shows, this indicates a strong positive correlation in the data between the final offers and the incidence of employer victories. The remarkable result in column (1), however, is the extraordinarily close match between the independent estimates of the mean and standard deviation of arbitrator preferences from the conventional and final-offer arbitration cases. This indicates that the coefficient of $(w^e + w^u)/2$ in the probit function (4) is nearly identical to the reciprocal of the standard deviation of the regression errors in equation (10). As Table 1 indicates, precisely the same result holds in the 1978 and 1979 data. Although not conclusive, this strongly suggests that the settlements in the conventionally arbitrated cases mainly reveal the underlying distribution of arbitrator preferences. The results in the columns (2) are obtained by maximizing the likelihood function (15) in order to test the joint hypothesis $\mu^r = \mu^p$ and $\sigma^r = \sigma^p$. As expected, a likelihood ratio test indicates that this hypothesis cannot be rejected in any of the years 1978, 1979, or 1980.

There is another simple way to state these empirical results. Suppose that we take the mean and standard deviation of the conventional arbitration awards as measures of the mean and standard deviation of arbitrators' preferences (w^a). These assumptions imply that we may predict the actual fraction of employer victories from knowledge of the final offers. Table 2 confirms that these predictions are essentially identical to the maximum-likelihood predictions based on the final-offer data alone. Likewise, these same assumptions imply that the correlation between the simple average of the union and employer final offers, $(w^e + w^u)/2$, and the incidence of employer victories may be predicted from the mean and standard deviation of the conventional arbitration awards alone. We also find this to be the case from the results in Table 2.

In the columns (3) in Table 2, we continue to maintain the hypothesis $\mu^p = \mu^r$ and $\sigma^p = \sigma^r$, and test the hypothesis $\alpha = 1.0$. In 1980, the estimate of α is well above unity, while in 1979 the estimate of α is slightly below unity, but in no case can we reject the hypothesis $\alpha = 1.0$. In the columns (4), we no longer maintain the hypotheses $\mu^p = \mu^r$ and $\sigma^p = \sigma^r$ and again test the hypothesis $\alpha = 1.0$. Again for 1979 and 1980 we can find no evidence for rejection of this hypothesis. This suggests that the unbiased decision rule (1) provides a satisfactory fit to the data, at least for 1979 and 1980. The exception is for 1978, where the hypothesis $\alpha = 1.0$ may be rejected at any conventional test level. What this finding represents in the data may be stated in a different way. In particular, the results in column (4) of Table 2 for 1978, for example, imply a coefficient (and standard error) for w^e in the probit equation (6) of .59 (.07). The estimated coefficient (and standard error) on w^u in equation (6) is $-.15$ (.04), however. These results indicate that the incidence of employer victories was positively correlated with the size of employer final offers in the 1978 data, as expected, but that the incidence of employer victories was negatively correlated with the size of union final offers, which was not expected. In view of the results for 1979 and 1980, we are inclined to attribute the poor performance of the simple arbitrator rule (1) in 1978 to the considerable confusion surrounding the initial implementation of the New Jersey statute.[12] Why and how this would have produced the anomalous empirical results in Table 2 is an issue that requires further research.

B. *Further Results*

Table 3 contains unrestricted estimates of the parameters β^p and β^r under various specifications for the variables in the vector x_i. Definitions and descriptive statistics for these variables are contained in Table 4. Also listed in each column are the values for the maximized logarithmic likelihood under the constraints $\beta^p = \beta^r$ and the maximized logarithmic likelihood under the further constraint that $\alpha = 1.0$.[13] Testing the hypothesis $\alpha = 1.0$ without maintaining $\beta^p = \beta^r$ may be accomplished by using a straightforward normal statistic computed from the estimate of α and its estimated standard error in Table 3.

The rationale for including the variables *SAL*, *TAX*, and *DEBT* in the analysis has already been established, but the inclusion of

---

[12] We initially advanced this explanation for the anomalous findings for 1978 in an earlier draft of this paper that contained data for 1978 and 1979 only. Our findings from the 1980 data obviously reinforced the strength of our belief in this explanation, but only additional data can confirm it. There was considerable (perhaps unnecessary) confusion at the outset of the implementation of the New Jersey statute because of an ambiguity in the CAP law then in effect. This law sets a limit on the growth of municipal expenditures and there was initial uncertainty regarding whether this was to be applied uniformly to each budget item, including police salaries. Employers argued that the CAP law did apply to each budget item and that arbitrators were bound by it also. Subsequent court decisions made it clear that this was not the case.

[13] The fitted coefficients for the constrained results are not reported to conserve space, but they are available from the authors in the form of a photocopied appendix.

TABLE 3—ESTIMATES OF EQUATIONS (6′) AND (11), THE DETERMINANTS OF THE PROBABILITY OF EMPLOYER VICTORIES, POLICE ARBITRATION CASES, NEW JERSEY, 1978–80[a]

Parameter	Estimated Models 1980 Cases (1)	(2)	(3)	Estimated Models 1979 Cases (1)	(2)	(3)	Estimated Models 1978 Cases (1)	(2)	(3)
α	1.369	0.932	1.007	0.658	0.817	0.832	−3.831	−6.454	−2.834
	(.238)	(.206)	(.224)	(.121)	(.103)	(.103)	(1.019)	(1.585)	(.503)
β^p(constant)	5.523	84.960	121.517	−10.749	−31.273	−81.843	7.762	53.126	22.352
	(.563)	(18.258)	(17.952)	(.936)	(3.029)	(7.687)	(.532)	(7.511)	(1.948)
$\beta^p(\Delta PCI_{-1})$	0.211	0.238	0.434	1.806	1.701	1.176	−0.269	−0.204	−1.432
	(.044)	(.047)	(.101)	(.103)	(.186)	(.134)	(.055)	(.047)	(.193)
$\beta^p(\text{Log}\,SAL_{-1}\times 100)$		−0.081	−0.125		0.022	0.081		−0.047	−0.0014
		(.018)	(.019)		(.003)	(.008)		(.008)	(.0003)
$\beta^p(TAX_{-1})$			0.023			0.001			0.014
			(.004)			(.001)			(.002)
$\beta^p(DEBT_{-1})$			−0.006			−0.006			−0.003
			(.001)			(.001)			(.001)
σ^p	1.518	1.233	1.067	2.067	2.031	1.387	2.368	1.933	1.029
	(.284)	(.158)	(.168)	(.363)	(.322)	(.158)	(.375)	(.337)	(.157)
β^r(constant)	0.878	31.950	21.287	2.976	−142.071	−1.334	5.182	22.914	11.061
	(.925)	(20.345)	(17.030)	(4.353)	(38.984)	(11.044)	(1.250)	(22.040)	(4.796)
$\beta^r(\Delta PCI_{-1})$	0.593	0.534	0.659	0.511	0.944	1.008	0.187	0.175	−0.398
	(.072)	(.080)	(.077)	(.394)	(.326)	(.127)	(.166)	(.166)	(.215)
$\beta^r(\text{Log}\,SAL_{-1}\times 100)$		−0.031	−0.020		0.145	0.015		−0.018	−0.003
		(.020)	(.017)		(.039)	(.012)		(.023)	(.005)
$\beta^r(TAX_{-1})$			−0.008			−0.044			−0.007
			(.002)			(.003)			(.002)
$\beta^r(DEBT_{-1})$			0.002			−0.011			−0.007
			(.001)			(.001)			(.002)
σ^r	1.194	1.153	0.952	2.176	1.671	0.310	2.181	2.166	1.910
	(.149)	(.144)	(.119)	(.344)	(.264)	(.049)	(.225)	(.223)	(.197)
Log Likelihood	−97.19	−87.62	−66.94	−80.82	−75.08	−32.22	−170.27	−167.70	−142.79
Log Likelihood for $[\beta^p,\sigma^p]=[\beta^r,\sigma^r]$	−100.16	−90.43	−89.85	−83.50	−80.08	−63.66	−171.69	−169.30	−161.02
Log Likelihood for $[\beta^p,\sigma^p]=[\beta^r,\sigma^r]$ and $\alpha=1.0$	−100.23	−90.43	−89.85	−83.50	−80.09	−64.63	−176.09	−173.13	−162.06

[a] Estimated standard errors are shown in parentheses.

TABLE 4—DEFINITIONS, MEANS, AND STANDARD DEVIATIONS OF THE CONTROL VARIABLES USED TO COMPUTE RESULTS IN TABLE 3

Variable	Definition	1980 FOA	1980 CA	1979 FOA	1979 CA	1978 FOA	1978 CA
ΔPCI_{-1}	Lagged percent change in county income per capita	12.49 (1.57)	12.44 (2.96)	10.75 (.77)	10.98 (1.26)	7.49 (.79)	7.30 (1.93)
SAL_{-1}	Salary level in previous year (in dollars)	18,136 (2248)	17,905 (2095)	17,091 (2529)	15,608 (1622)	16,227 (1966)	16,278 (2292)
TAX_{-1}	Municipality's state equalized tax rate in previous year (in dollars per 100 dollars of equalized value)	2.58 (.66)	2.63 (.76)	2.90 (.99)	3.11 (.56)	3.07 (1.17)	4.11 (1.79)
$DEBT_{-1}$	Net debt per capita in municipality in previous year (in dollars)	209 (143)	250 (174)	161 (101)	161 (170)	190 (127)	216 (163)

Note: See Table 2.

ΔPCI, the percentage change in county income per capita, requires some explanation. As our earlier discussion suggested, we would like to have measured the percentage change in some alternative wage rate in the municipality to include in the vector of variables x_i. At the level of the municipality to which these data refer, however, the best measure of this quantity that we could obtain was the percentage change in per capita income in the county of which the municipality is a part. The use of this variable no doubt entails some measurement error and may result in downward-biased estimates of the effect of this variable. A useful future research project would be to remedy this measurement defect, if this is possible.

As Table 3 indicates, the alternative wage variable (ΔPCI) has a statistically significant and positive effect on the mean of arbitrator preferences in the data for 1979 and 1980 for both the final-offer and conventional arbitration cases. In 1979, the alternative wage coefficient is close to unity for both types of cases, as we had speculated might be the case. In 1980, however, this coefficient is significantly less than unity, and in 1978, it is often negative and occasionally statistically significant. Taken together, these results provide some support for the hypothesized importance of alternative wages in the determination of arbitrator preferences, although the results are far from conclusive.

In 1978 and 1980, the level of the previous contract salary has a negative effect on arbitrator preferences that is generally statistically significant. This suggests that in these years arbitrators tended to behave as if they wished to narrow the salary differences across municipalities. In 1979, however, the previous contract salary tends to have a positive effect on arbitrator preferences that is often statistically significant. Taken at face value, these results suggest that the arbitrator's perceived role in narrowing (or widening) intermunicipality salary differences may be unstable from year to year.

Finally, in 1978 and 1979, greater levels of taxation and per capita municipal debt have an unambiguously negative and statistically significant effect on arbitrator awards. In 1980, these results are no longer unambiguous; indeed, the (unreported) estimates that constraint $\beta^p = \beta^r$ generally show no statistically significant effect of per capita debt or taxation on arbitrator awards. These results suggest that the arbitrator's perceived role in responding to municipal taxation or debt burdens is by no means stable from year to year.

In sum, we have found statistically significant, but temporally unstable, effects of alternative wage rates, prior contract wage rates, and per capita debt and taxation variables on arbitrator awards. There are two alternative explanations for these results, but it will take additional research to sort out their empirical validity. First, it may simply be the case that the main predictable variability in the central tendency of arbitrator awards is in their movements through time. If this is the case, then our time-series of estimates of $\mu = \mu^p = \mu^r$ and $\sigma = \sigma^p = \sigma^r$ for each year should ultimately provide the data for a more complete explanation of the basic variability in arbitrator preferences. Considerably more experience with the New Jersey statute will be necessary before there are enough data to make this feasible. Moreover, if this is the explanation for our findings, then most of the parameter estimates in Table 3 do not reveal much, if anything, about the more fundamental determinants of arbitrator decision making. Indeed, in casual conversations we have found that much arbitrator behavior is characterized as "following the herd," or "looking up and down the street," and the results in Table 3 may simply be a description of this behavior rather than an explanation of it.

An alternative explanation is that there are stable relationships between arbitrator preferences and some set of variables x_i, but that in our analysis important variables have been omitted from x_i. Only data from a better experiment than we have available can confirm or refute this possibility.

The remaining results in Table 3 confirm the major findings in Table 2. In 1979 and 1980, the hypothesis $\alpha = 1.0$ is clearly accepted by the data, although for 1978 it clearly is rejected. On the other hand, the joint constraints $\beta^p = \beta^r$ and $\sigma^p = \sigma^r$ are occasionally rejected in the various specifi-

Table 5—Estimates of σ^P in Equation (4) with $\alpha = 1$ and with (and without) Arbitrator Dummies, Police Arbitration Cases, New Jersey, 1978–80[a]

	1980		1979		1978	
σ^P	4.885	2.870	3.185	3.156	2.272	1.114
	(3.153)	(1.607)	(1.532)	(1.514)	(.821)	(.294)
Log Likelihood	−34.89	−28.89	−21.82	−20.93	−36.89	−30.04
Arbitrator Dummies Included (No. included)	No (–)	Yes (3)	No (–)	Yes (3)	No (–)	Yes (6)
Number of FOA Observations	53[b]	53[b]	39	39	65	65

[a] Estimated standard errors are shown in parentheses.

[b] These data include several arbitration cases involving county police.

cations for x_i that are tested in Table 3. We are inclined to attribute this to the very small samples (especially for conventional arbitration) available for estimating the fully unconstrained models. No doubt larger samples and a more detailed set of variables for the vector x_i would increase the power and confidence with which these hypotheses can be tested.

Finally, and perhaps most important, the addition of the variables x_i to the analysis reduces the estimated unpredictable variability in arbitrator preferences (σ). Nevertheless, considerable unpredictable variability remains in all cases. In view of the importance of this issue, we have also pursued it by an attempt to account for the individual influence of specific arbitrators. To do this, we have estimated equation (4) with separate intercepts for each arbitrator. In the probit model, these fixed effects are econometrically identified only if an individual arbitrator selected at least one employer offer and one union offer in a given year. Table 5 lists the number of final-offer arbitration cases that satisfy this criterion. There are simply too few conventional arbitration cases to perform the analysis for that group. Depending on the year, the table indicates that nearly one-half of the final-offer arbitration cases are decided by four to seven arbitrators who selected at least one union and one employer final offer.

The results in Table 5 indicate that the introduction of the arbitrator dummies considerably reduces the estimate of the unpredictable variability (σ) in arbitrator preferences in 1980 and 1978. Nevertheless, σ^P never falls below 1 percent and is sometimes as high as 3 percent. This strongly suggests that successful models of the strategic behavior of the parties will almost certainly have to account explicitly for the uncertainty the parties face in arbitrator decision making.

III. Conclusion

The empirical models of arbitrator behavior in this paper are an important first step in the much-needed empirical studies of arbitration outcomes that remain on the agenda for further research. More complete models should explain both the determination of the parties' final offers, whether they choose to engage in conventional or final-offer arbitration, or whether they choose to negotiate a settlement without arbitration. The simple models of arbitrator behavior presented and estimated with remarkable success here will presumably be a key ingredient to this further research.

We have two important substantive empirical results to report. First, under the New Jersey statute, union offers have been selected most of the time in final-offer arbitration cases.[14] This finding has raised immediate questions about the impartiality and integrity of the arbitrators in some quarters. In general, we should expect a fair arbitrator to be one that considered the objective considerations in a particular case and then settled on what, in the arbitrator's own mind, seemed a preferred settlement. Given the arbitrator's preferred award, it seems clear that a fair arbitrator must select whichever final offer is closest to it. Supposing that the union and employer also understand this process, they will use their best estimates of the arbitrator's preference to shape their own offers with the understanding that a higher offer by either party will increase the probability that the employer's offer will be selected. If the parties behave symmetrically, as most of us might have expected, we would expect the union and employer offers to fall equally distant from, but on opposite sides of, the parties' best estimate of the arbitrator's preferred award. In this scenario, we should naturally expect the union offers to be selected in one-half of the cases. It is this prediction that is strongly contradicted by the facts.

---

[14] This phenomenon is not restricted to New Jersey, however. There is evidence of similar behavior in Michigan and Massachusetts, as noted in our 1983 paper.

It follows that there are two different reasons why union offers may be disproportionately accepted by the arbitrators. On the one hand, arbitrators may not follow the decision process set out above. If this is the case, the integrity of the arbitration system is being seriously undermined. On the other hand, the parties may not typically position themselves equally distant from, and on opposite sides of, the arbitrator's preferred award. This might happen either because unions have a more conservative view of what arbitrators will allow, or because unions may be more fearful of taking the risk of loss than are employers.

Our economic results provide strong evidence, however, that the final-offer arbitration decisions were generated by a set of impartial arbitrators who were systematically applying the same standards used in conventional arbitration cases. The evidence for this is that (a) the proportion of union victories and (b) the correlation between the incidence of union victories and the mean of the union and employer final offers are precisely those predicted from the data on conventional arbitration cases alone. (These are econometric tests of standard cross-equation restrictions.) These results indicate that arbitrators treat generous employer offers no differently than they treat conservative union offers. Instead, the union offers are most often selected because the frequency of conservative union offers is considerably greater than the frequency of generous employer offers. It follows that any critique of the generosity of the final-offer arbitration awards must be based on something other than simple win-loss records.

Our second important empirical result is that this conservative union behavior resulted in *lower* average wage increases under the final-offer arbitration provisions than under the conventional arbitration provisions of the New Jersey statute. Although the union offers are accepted in a vast majority of the final-offer arbitration cases, and so the unions appear to win in final-offer arbitration, the actual wage increases in these cases are lower than in the conventional arbitration cases. At the same time, we find considerably less variability in the actual awards under the final-offer arbitration provisions than under the conventional arbitration provisions of the statute. What the union bargainers gave up, therefore, by way of a decrease in the mean award under final-offer arbitration, they may have made up by a reduction in its variability. This suggests that union bargainers may be considerably more risk averse than employer bargainers and opens up some interesting possibilities for additional research.

The serious quantitative study of dispute settlement mechanisms has only just begun. Further empirical and theoretical research on the operation of these mechanisms may therefore have an important impact on their future design.

REFERENCES

Ashenfelter, Orley, "Determining Participation in Income-Tested Social Programs," *Journal of the American Statistical Association*, September 1983, *78*, 517–25.

______ **and Bloom, David E.,** "The Pitfalls in Judging Arbitrator Impartiality by Win-Loss Tallies under Final-Offer Arbitration," *Labor Law Journal*, August 1983, *34*, 534–39.

Bloom, David E., "Customized 'Final-Offer': New Jersey's Arbitration Law," *Monthly Labor Review*, September 1980, *103*, 30–33.

______, "Is Arbitration *Really* Compatible with Bargaining?," *Industrial Relations*, Fall 1981, *20*, 233–44.

Card, David, "Arbitrators as Lie Detectors: A Suggested Interpretation," unpublished paper, Graduate School of Business, University of Chicago, January 1983.

Crawford, Vincent P., "On Compulsory-Arbitration Schemes," *Journal of Political Economy*, February 1979, *87*, 131–60.

Davis, Otto A., "Public and Private Characteristics of a Legal Process: A Comment," *Journal of Legal Studies*, March 1979, *8*, 285–93.

Farber, Henry S., "Mechanisms for Settling Public Sector Labor Disputes: A Comparative Evaluation of Conventional Arbitration and Final-Offer Arbitration," unpublished paper, August 1979.

______, "An Analysis of Final-Offer Arbitration," *Journal of Conflict Resolution*, December 1980, *5*, 683–705.

______, "Splitting-the-Difference in Interest Arbitration," *Industrial and Labor Relations Review*, October 1981, *35*, 70–77.

______ **and Katz, Harry C.**, "Interest Arbitration, Outcomes, and the Incentive to Bargain: The Role of Risk Preferences," *Industrial and Labor Relations Review*, October 1979, *33*, 55–63.

Gould, John, "The Economics of Legal Conflicts," *Journal of Legal Studies*, June 1973, *2*, 279–300.

Landes, William and Posner, Richard, "Adjudication as a Private Good," *Journal of Legal Studies*, March 1979, *8*, 235–84.

Stevens, Carl M., "Is Compulsory Arbitration Compatible with Bargaining?," *Industrial Relations*, February 1966, *5*, 38–52.

The Economics of Price Scissors

By RAAJ KUMAR SAH AND JOSEPH E. STIGLITZ*

An important policy issue facing many developing countries is that of the appropriate terms of trade between the industrial and agricultural sectors. Traditionally, this problem has been known as the "scissors problem," and it has been typically discussed in the context of a socialist society striving for capital accumulation in the early phase of economic development.[1] Changing the terms of trade has distributional consequences as well. We examine here both the accumulation and the distribution aspects simultaneously.

Our analysis is based upon a simple general equilibrium model of a dual economy, in which the role of incentives in the rural sector is emphasized. Using this model, we describe the effects of altering the terms of trade on the members of the rural and urban sectors, and on the state's surplus. We then extend our analysis to obtain normative results, such as: under what conditions will a change in the terms of trade unequivocally benefit a society; and what is the nature of the optimal terms of trade?

The issues being examined in the present paper are of considerable historical as well as contemporary significance. One of the central concerns of the classical economists was the relative roles of the town and the country in the early phases of economic growth. Probably the best known landmark of this interest was the lively controversy between Thomas Malthus and David Ricardo on the corn laws. These concerns were also recognized, but side-stepped, by Karl Marx.[2]

The town vs. country questions emerged as pivotal, however, on the eve of the October Revolution; so much so that every important Soviet leader had to grapple with it.[3] Furthermore, the economic events of the early 1920's were so severe (see Maurice Dobb, ch. 7) that the state's policy on the terms of trade became a raging controversy in the ensuing debate on Soviet industrialization. A seminal participant in this debate was Evgeny Preobrazhensky, who proposed that the state can, and should, increase its surplus by turning the terms of trade against peasants.[4] This policy of "primitive accumulation" was challenged on a number of different grounds by many members of the Soviet leadership (see Alexander Erlich, 1960). What interests us here about the Soviet debate is that some of our results clarify and correct certain crucial elements of this debate.

Even though the discussions within the USSR itself had subsided with the institution of Joseph Stalin's collectivization policy, the town vs. country problems have remained important in recent decades in many socialist countries, for example, in Eastern Europe

---

*Departments of Economics, University of Pennsylvania, Philadelphia, PA 19140 and Princeton University, Princeton, NJ 08544, respectively. We thank Jere Behrman, Michael Lipton, and an anonymous referee for comments on an earlier draft of this paper. Stiglitz's research was supported by the National Science Foundation. Stiglitz would like to thank A. Braverman for helpful discussions.

[1] The term price scissors was used extensively in the Soviet industrialization debate. It denotes the relative price of the industrial (urban) good in terms of the agricultural (rural) good. It is interesting to note that the diagram of the relative movements of the retail prices of agricultural and industrial goods in the USSR, between April 1922 and March 1924, indeed resembles a pair of scissors (Maurice Dobb, 1966, p. 164).

[2] "The Foundation of every division of labour...is the separation of town from country. One might well say that the whole economic history of society is summed up in the movement of this antithesis. However, for the moment we shall not go into this" (1967, p. 472). Marx does not elaborate on this issue in his later writing either.

[3] For example, Nikolai Bukharin (1971), Vladimir Lenin (1975a; b), Evgeny Preobrazhensky (1965), Joseph Stalin (1954), and Leon Trotsky (1971).

[4] Preobrazhensky's verbal model consisted of a peasant sector existing side by side with a state controlled industrial sector. This model, as Avinash Dixit correctly points out (1973, p. 325), can be considered a precursor of the modern models of dual developing economies.

and in the People's Republic of China. Specifically, the past attempts of several socialist economies to achieve direct control of the agricultural sector have not been particularly successful, and these experiences in turn have added greater relevance to the examination of indirect control mechanisms, as, for example, through price incentives.

Similar considerations are relevant in developing economies. In fact, in a typical developing economy, not only does the government have fewer instruments of control, but also the agricultural sector plays a more important role in determining national output and employment. It is not surprising, therefore, that the issue of intersectoral pricing is a central aspect of policymaking in many developing economies.[5] The research on these issues, however, has been surprisingly inadequate.[6] In particular, an analytical examination of the scissors problem has been hitherto lacking, and the present paper is intended to fill this lacuna.

The plan of the paper is as follows. The basic model of the economy is presented in Section I. Descriptive analysis is conducted in Sections II and III. The analysis of normative questions is then taken up. Results on the desirable reforms in the terms of trade are presented in Section IV, and the optimal tax and the optimal terms of trade are discussed in Section V. The model is then expanded in Section VI to incorporate labor mobility and unemployment. Section VII contains some comments on the Soviet industrialization debate. Certain extensions of the present analysis are discussed in Section VIII.

I. The Model

The rural sector's population is N^1, and A is the total agricultural land area which is owned within the rural sector. Intrasectoral distribution is ignored, and $a = A/N^1$ is the land area per worker. The term L^1 denotes the hours worked by each worker. The production technology exhibits constant returns to scale, and $X \equiv X(A/N^1, L^1) \equiv X(a, L^1)$ is the output per worker. A rural worker's consumption of the rural and the urban good is denoted by (x^1, y^1). The surplus of the rural good per worker, Q, is given by

$$(1) \qquad Q = X - x^1.$$

The relative price of the rural good in terms of the urban good is denoted by p. A rural worker's budget constraint becomes

$$(2) \qquad pQ = y^1.$$

A rural worker's utility is represented by $U^1 = U(x^1, y^1, L^1)$. The indirect utility is obtained from

$$(3) \quad V^1(p, N^1) = \max_{x^1, y^1, L^1} U(x^1, y^1, L^1) + \lambda^1[pX(A/N^1, L^1) - px^1 - y^1].$$

From the envelope theorem

$$(4) \qquad \partial V^1/\partial p = \lambda^1 Q > 0,$$

$$\text{and} \qquad \partial V^1/\partial N^1 = -\lambda^1 p X_a a/N^1 < 0,$$

where $X_a = \partial X/\partial a$, and λ^i is the (positive) marginal utility of income in sector i. It is evident that the rural surplus is a function of the relative price and the rural population, that is, $Q \equiv Q(p, N^1)$. It is also worth keeping in mind that, in our notation, an increase in the size of price scissors, that is, an increase in the relative price of the urban good, corresponds to a decrease in p.

The urban population is N^2, and an urban worker supplies L^2 hours of work which are fixed by the government, based on technological considerations. The urban worker's consumption of the rural and the urban good is denoted by (x^2, y^2), and w is the wage per hour, in terms of the industrial good. The budget constraint of an urban worker is given by

---

[5] This may also explain why there has been an increasing sense of dissatisfaction with the quantity-based approach of the traditional theories of economic planning. These approaches, originating in the model of G. A. Fel'dman (see Nicholas Spulber, 1964), typically ignore the rural sector entirely, which implicitly means that the government can force the rural sector to deliver the desired amount of food needed for the urban workers, *regardless* of the amount of urban goods the rural sector receives in return.

[6] A significant recent work in this area is by Michael Lipton (1977), in which a history of the town vs. country ideologies is traced. Also, see Dixit (1969), and J. M. Hornby (1968).

$$px^2 + y^2 = wL^2. \tag{5}$$

Denoting the urban individual's utility function by $U^2 = U(x^2, y^2, L^2)$, the indirect utility is obtained from[7]

$$V^2(p,w) = \max_{x^2, y^2} U(x^2, y^2, L^2) + \lambda^2\left[wL^2 - px^2 - y^2\right]. \tag{6}$$

The envelope theorem yields

$$\partial V^2/\partial w = \lambda^2 L^2 > 0, \tag{7}$$

and

$$\partial V^2/\partial p = -\lambda^2 x^2 < 0.$$

Naturally, the urban consumption is a function of the relative price and the urban wage, that is, $x^2 \equiv x^2(p,w)$. The consumption goods are assumed to be normal. The output of an urban worker is denoted by $Y \equiv Y(k, L^2)$, where $k = K/N^2$ is capital stock per urban worker, and K is the total urban capital stock.

The economy under consideration is a closed economy.[8] The sectoral populations are fixed. The total population is N, and

$$N = N^1 + N^2. \tag{8}$$

The rural quantities cannot be controlled directly, but they can be influenced indirectly through the terms of trade, p. The urban sector is somewhat more controllable, in that the urban wage can be changed by the government.[9]

The two basic constraints in the economy are the quantity balances of the urban and rural goods, respectively. Defining I as the state's surplus of the urban good, we have

$$I = N^2Y - N^2y^2 - N^1y^1; \; I = \dot{K}. \tag{9}$$

That is, the urban output is used either for consumption or for investment. Similarly, the balance between the supply and demand of the rural good requires

$$N^1Q(p, N^1) = N^2x^2(p,w). \tag{10}$$

Finally, note that the above model can be easily expanded to include the possibility of investing the urban capital good in the rural sector. We ignore this aspect because the focus of the present paper is on the analysis of the terms of trade. This analysis in any case remains essentially unchanged, as we will see later, if the investment allocation decision is incorporated.

With the features of the economy described above, we can now begin the analysis of the terms of trade. The next two sections trace the impact of altering the terms of trade on the basic economic variables. Interestingly enough, six decades ago, Preobrazhensky had presented certain important propositions on these matters. Our descriptive analysis ascertains the conditions under which Preobrazhensky's propositions are valid.

II. Preobrazhensky's First Proposition

Preobrazhensky's most important claim was that *the state can increase accumulation by moving the terms of trade against peasants, that is, by increasing the size of scissors*. For ease of reference, we call this Preobrazhensky's first proposition. Naturally, this proposition is quite basic, because it asserts the feasibility of using the terms of trade as an instrument for society's accumulation.

Recall that moving the terms of trade against peasants means a decrease in p. The above proposition thus says: $dI/dp < 0$. To examine the conditions under which this proposition holds, we obtain an alternative expression for the state's surplus. Substitution of (2), (5), and (10) in (9) yields

---

[7] Individuals' preferences are assumed identical in the two sectors, but the more general case can be easily worked out.

[8] As we discuss later, this is the correct assumption to make for examining the scissors problem. For an analysis of pricing in an open economy, see our 1983b paper.

[9] We discuss other possible instruments of policy, as well as other structures of the economy, in Section VIII.

$$I = N^2(Y - wL^2). \tag{11}$$

The state's surplus is therefore the difference between the urban output and the wage payment to the urban workers, and the surplus does not directly depend on the terms of trade, p.

A change in p, however, requires a change in the urban wage w if the demand and supply balance of the rural good, (10), is to be preserved. To obtain the relation between p and w, we first define some notation. Let $m = wL^2$ be the income of an urban worker,

$$\varepsilon^1_{Qp} = \partial \ln Q / \partial \ln p, \tag{12}$$

$$\varepsilon^2_{xp} = -\partial \ln x^2 / \partial \ln p,$$

and $$\varepsilon^2_{xm} = \partial \ln x^2 / \partial \ln m$$

be, respectively, the price elasticity of the rural surplus, the price elasticity of the urban consumption of the rural good, and the income elasticity of the urban consumption of the rural good. Also, define ε_{wp} as the percent change in the urban wage which must accompany a percent increase in p. That is,

$$\varepsilon_{wp} = \partial \ln w / \partial \ln p. \tag{13}$$

Next, differentiation of (10), and the use of the above definitions yields

$$\varepsilon_{wp} = \left(\varepsilon^1_{Qp} + \varepsilon^2_{xp}\right) / \varepsilon^2_{xm}. \tag{14}$$

The change in the state's surplus due to a change in the terms of trade is then obtained from (11) and (14) as

$$dI/dp = -N^2 wL^2 \left(\varepsilon^1_{Qp} + \varepsilon^2_{xp}\right) / p\varepsilon^2_{xm}. \tag{15}$$

In the above expression, the sign of the rural surplus response, ε^1_{Qp}, is not predictable theoretically because this response is a composite of the production response, the consumption response, and the labor supply response. Fortunately, however, several sources of empirical evidence already exist on this subject. Among them are (i) analysis of rural surplus in the Soviet Union during the 1920's (Dobb, ch. 7), (ii) econometric studies of economywide supply response of different crops in many countries, following an early work by Jere Behrman (1968), and (iii) microeconometric studies of farm-household behavior (Lawrence Lau, Wuu-Long Lin, and Pan Yotopoulos, 1978, and Howard Barnum and Lyn Squire, 1979). All of these studies indicate a positive surplus response to price. Based on this evidence, we maintain throughout the paper that

$$\varepsilon^1_{Qp} > 0. \tag{16}$$

Now, looking at the right side of (15), ε^2_{xp} and ε^2_{xm} are positive because consumption goods are normal. Using (16), therefore, $dI/dp < 0$. (In fact, the same result will hold even if ε^1_{Qp} is negative within some range.) Further, note from (15) that a higher ε^1_{Qp} corresponds to a higher magnitude of dI/dp.

We therefore conclude that *Preobrazhensky's first proposition is valid*, *that is*, *turning the terms of trade against peasants leads to an increased accumulation*. Also that *turning the terms of trade against peasants leads to a larger increase in accumulation if the price response of the rural surplus is larger*.

The above results need to be contrasted with the assertion made by several authors that increasing the squeeze on farmers through the price scissors may not lead to accumulation if the rural surplus is highly price responsive (for example, Michael Lipton, pp. 129–30, and Ashok Mitra, 1977, p. 54). As shown above, not only is this not correct, but quite the opposite is true!

The main point missed by earlier researchers is the distinction between the rural surplus and the state's surplus, and the distinction between the state's surplus and the consumption of industrial workers. Surely, a decrease in the relative price of the rural good will lead to a lower rural surplus, if the price response of the rural surplus is higher. This in turn will require a higher decrease in the urban demand of the rural good, and the only way in which it can be accomplished is if a higher decrease in the urban real wage is enforced by the government. This higher decrease in the urban wage will then naturally lead to a higher state's surplus.

III. Preobrazhensky's Second Proposition

The economic content of another important claim of Preobrazhensky can be expressed as follows: *by turning the terms of trade against peasants, it is possible to accumulate in a manner that the economic position of industrial workers will not deteriorate.*[10]

This proposition can be expressed in our notation as: $dV^2/dp \leq 0$. From the earlier analysis, it is clear that there are two effects of changing the terms of trade on an urban worker: a direct price effect, and an indirect effect due to the induced change in the urban wage. The total effect is expressed as

$$\frac{dV^2}{dp} = \frac{\partial V^2}{\partial p} + \frac{\partial V^2}{\partial w}\frac{dw}{dp}. \tag{17}$$

Substitution of (7) in the above yields

$$\frac{dV^2}{dp} = \lambda^2\left[-x^2 + \frac{wL^2}{p}\varepsilon_{wp}\right]. \tag{18}$$

We define $\alpha_x^2 = px^2/wL^2$ as the urban budget share of the rural good, and $\varepsilon_{xp}^{2u} = -(p/x^2)(\partial x^2/\partial p)_{\bar{V}^2}$ as the price elasticity of the compensated demand for the rural good in the urban sector. Using these definitions, the Slutsky expression can be written as

$$\varepsilon_{xp}^2 = \varepsilon_{xp}^{2u} + \alpha_x^2\varepsilon_{xm}^2. \tag{19}$$

Substitution of (14), and (19), in (18) gives

$$\frac{dV^2}{dp} = \frac{\lambda^2 x^2}{\alpha_x^2}\frac{\left(\varepsilon_{Qp}^1 + \varepsilon_{xp}^{2u}\right)}{\varepsilon_{xm}^2}. \tag{20}$$

In the above expression, $\varepsilon_{xp}^{2u} \geq 0$ from the standard Slutsky property of compensated demand. It follows from (16) and (20) that $dV^2/dp > 0$. Also note that a higher ε_{Qp}^1 corresponds to a higer dV^2/dp.

Therefore: *the welfare of industrial workers must decline if the state accumulates by turning the terms of trade against peasants. Preobrazhensky's second proposition is therefore invalid.* Also: *turning the terms of trade against peasants leads to a larger decline in the welfare of industrial workers if the price response of the rural surplus is larger.*

The crucial point missed by Preobrazhensky and subsequent researchers is the constraint generated by the demand and supply balance of the rural good.[11] This constraint, (10), dictates the feasible combinations of the terms of trade and the urban wage and, hence, it determines the feasible (as well as the necessary) change in the urban wage corresponding to any change in the terms of trade.[12] This in turn determines the effect of a change in the terms of trade on the state's surplus as well as on the welfare of urban workers.

To recapitulate, we have shown that turning the terms of trade against the rural sector leads to an increase in the state's surplus according to (15), and that it leads to a decrease in the welfare of urban workers according to (20). Finally, the effect of the terms of trade on the rural sector's welfare is given directly by (4). As one would expect, the peasants experience a decrease in welfare if the price ratio is turned against them.

---

[10] "The objection that taxation on the basis of a definite price policy will affect the wages of workers... is completely futile.... Here is a numerical example: if as a result of an appropriate price policy the working class along with the rest of the population pays to the state industry say 50 million, the state can easily return this sum to the workers by an increase in wages,..." (Preobrazhensky, p. 112).

[11] It is in fact possible that some governments are also ignorant of such a constraint. If this is the case, then a government may decide to lower the relative price of food *without* changing the urban wage. A natural response to the resulting shortage of food in cities would be to introduce an urban rationing system. In turn, then, one could argue that the recurring occurrences of rationing in many economies is consistent with the kind of ignorance suggested above. In any event, a rationing system, whether it is introduced by design or because the government is unaware of the constraints in the economy, entails a different set of instruments than the one on which we focus in this paper. See our 1983b paper for further analysis.

[12] Contrast this to fn. 10, which appears to suggest that the urban wage can be arbitrarily changed along with any given change in the terms of trade.

These three effects are now put together for normative analysis, which is conducted in the next two sections.

IV. Reform in the Terms of Trade

The aim of a policy reform analysis is to identify rules for an improvement in society's overall welfare. Naturally any such rule is more useful if less information is required to apply it. The results in this section are particularly significant from this point of view because, as we shall see, their application requires quite minimal information.

The first step in the analysis of reform is to define the aggregate social welfare. For this, we use an additive Bergson-Samuelson social welfare function, ψ, to aggregate over individual utilities.

$$(21)\quad \psi = N^1W(V^1)+N^2W(V^2),$$

where W is concave and increasing in V. If δ is the social value of the marginal investment, then the current value of the aggregate social welfare, H, is given by

$$(22)\quad H=\psi+\delta I.$$

Substituting (11) in (22)

$$(23)\quad H=N^1W(V^1(p,N^1)) + N^2W(V^2(p,w))+\delta N^2[Y-wL^2].$$

The above expression is now perturbed with respect to p, while ensuring a corresponding perturbation in the urban wage to preserve the identity (10). Using (4) and (7), we obtain

$$(24)\quad \frac{dH}{dp}=\frac{\partial H}{\partial p}+\frac{\partial H}{\partial w}\frac{dw}{dp} = N^1\beta^1Q+N^2\beta^2\left[-x^2+\frac{wL^2}{p}\varepsilon_{wp}\right] - \delta N^2wL^2\varepsilon_{wp}/p,$$

where $\beta^i=\lambda^i\partial W/\partial V^i$ is the social value of a marginal increase in the income of a worker in sector i. The three terms on the right-hand side of (24) represent, respectively, the welfare gain to the rural sector, the welfare gain to the urban sector, and the loss of investment due to an increase in the relative price of the rural good. Naturally, these gains and losses are weighed by their respective social weights.

Using (10), and the definition of the budget share, α_x^2, (24) can be expressed as

$$(25)\quad dH/dp=N^2x^2(\beta^1-\beta^2) + N^2x^2(\beta^2-\delta)\varepsilon_{wp}/\alpha_x^2.$$

Note that the above expression clearly separates the two distinct effects of a change in the relative price. The first term is the direct effect of an increase in the relative price of the rural good, which benefits the peasants but hurts the urban workers. The second term is the net effect of the induced increase in the urban wage, which benefits the urban workers but reduces the investment fund.

The expression (25) can now be used to obtain the rules for reform in the terms of trade, that is, the sufficient conditions which will guarantee that a specific change in the relative price will increase the social welfare. Rewriting (25), we obtain

$$(26)\quad dH/dp=N^2x^2\left[(\beta^1-\delta) + (\beta^2-\delta)\left(\varepsilon_{wp}/\alpha_x^2-1\right)\right].$$

Next note from (19) that $\varepsilon_{xp}^2/\alpha_x^2\varepsilon_{xm}^2\geq 1$, because $\varepsilon_{xp}^{2u}\geq 0$. Further, as $\varepsilon_{Qp}^1>0$, it follows from (14) that

$$(27)\quad \varepsilon_{wp}/\alpha_x^2>1.$$

Substitution of the above in (26) yields the following two rules for reforms:

$$(28)\quad dH/dp<0,\ \text{if } \beta^1\leq\delta \text{ and } \beta^2\leq\delta$$
(with at least one strict inequality);

$$(29)\quad dH/dp>0,\ \text{if } \beta^2\geq\delta \text{ and } \beta^2\geq\delta$$
(with at least one strict inequality).

From (28) and (29), we find that: *moving the terms of trade against (in favor of) peasants is desirable if the social weight on investment is greater (smaller) than the social weights on rural and urban income.* Note that this rule holds no matter which one of the two social weights on income is larger. The reason for this is quite simple. If the social valuation of a dollar of investment exceeds the social valuation of a dollar of consumption (in both sectors), then the gains from moving the prices against peasants exceed the losses due to decreased consumption, regardless of which one of the two sets of workers is worse off.

What is important about these rules is that their use does not require the knowledge of the behavioral parameters (such as the rural surplus response, and the urban consumption response). Our rules can be used solely on the basis of the social weights in the existing regime, that is, the social weights associated with a dollar of rural and urban income versus a dollar of investment. The simplicity of these rules is quite unlike the results which one typically finds in the public finance literature (see Anthony Atkinson and Stiglitz, 1980, pp. 382–86, for a recent review of this material).

The procedure for moving the terms of trade in the desirable direction can now be examined. Suppose we are in an initial regime in which the social weight on investment exceeds the two social weights on income. Then, from (28) the relative price of the rural good, p, should be decreased. A decrease in p, from (4) and (20), will decrease the individual utilities, V^1 and V^2, but, on the other hand, it will increase the investment I, due to (15). The social weights on income, β^1 and β^2, will in turn increase since individuals now have lower levels of utility.[13] In contrast, it is natural to expect that the decrease in p will lead to a decrease in the social weight on investment, δ, because the investment has become larger.

Thus, after every small decrease in the relative price, we need to compare β^1, β^2, and δ. The price decrease should be continued until that point when the larger of the two social weights on incomes becomes equal to the social weight on investment. Once this equality arises, any finite decrease in p will violate (28). Thus, we can no longer use the rules of reform, and additional information is needed to facilitate further policymaking. Essentially the same line of reasoning applies for an increase in the relative price if we are in an initial regime in which the social weight on investment is smaller than both of the social weights on income, that is, when (29) holds.

A final point needs to be emphasized here. The above analysis clearly shows that, over a range of social weights, the decision to widen or to narrow the price scissors is independent of the comparison between rural and urban welfare. Instead, this decision depends on the intertemporal comparison between consumption and investment.

V. Optimal Terms of Trade

We are already in a position to recognize some features of the optimal price structure. Assuming that a unique interior maximum exists for our maximization problem, we should have $dH/dp = 0$ at the optimum. The necessary conditions for this characteristic of the optimum can be extracted from (28) and (29). We find that the optimal price structure must satisfy[14]

$$(30) \qquad \beta^1 > \delta > \beta^2, \text{ or } \beta^2 > \delta > \beta^1.$$

In other words, if the social weight on investment does not lie between the social weights on the rural and the urban income, then the regime is not optimal, and it can be improved through price policy. This result needs to be contrasted with a view often held in the literature on project evaluation that the social weight on investment should be higher than the (undifferentiated) social weight on consumption (Yotopoulos and Jeffrey Nugent, 1976, p. 385, for example). The view is incorrect if the terms of trade can be used as an instrument of policy.

---

[13] This is because $\partial\beta/\partial V < 0$. To ensure this, we are assuming that either (i) the social welfare function is strictly concave and the marginal utility of income is nonincreasing in income, or (ii) the social welfare function is concave and the marginal utility of income is decreasing in income.

[14] To keep our exposition uncluttered, we are ruling out a rather rare regime in which $\beta^1 = \beta^2 = \delta$.

Before analyzing the optimal term of trade, we take up the question of taxation in the present policy problem. The recent literature on optimal tax theory suggests that more insights can often be obtained by comparing market prices with shadow prices, rather than by comparing producer and consumer prices (Stiglitz and Partha Dasgupta, 1971). This is especially relevant here since producer and consumer prices are identical. In fact, a rural worker is simultaneously a producer and a consumer.

We therefore define "tax" ("subsidy") as the difference between the consumer (producer) price and the social opportunity cost of producing the same good. If η is the shadow price of the rural good, then the shadow price of the rural good in terms of the urban good is η/δ. The subsidy rate therefore is defined as $s=(p-\eta/\delta)/p$. An interesting feature of the present problem is that the same price regime will mean a subsidy to one sector and a tax on another sector. For example, if $s>0$, then the urban sector is paying a tax while the rural sector is receiving a subsidy.

To obtain the optimal tax formula, the following Lagrangian is defined by explicitly incorporating the rural good constraint, (10), into the maximand

$$(31)\quad Z=\psi+\delta\left[N^2Y-N^2y^2-N^1y^1\right]+\eta\left[N^1Q-N^2x^2\right],$$

where ψ is given by (21), and p and w are the control variables.[15] The first-order conditions with respect to p and w, after some manipulation,[16] can be written as

$$(32)\quad s=\left(\beta^1-\beta^2\right)/\delta\left(\varepsilon^1_{Qp}+\varepsilon^2_{xp}\right),$$

$$(33)\quad s=\left(\delta-\beta^2\right)/\delta\alpha_x^2\varepsilon^2_{xm}.$$

Equations (32) and (33) are alternative expressions for the optimal subsidy rate, and these can be easily interpreted. Note from (32) that s is positive or negative depending on whether β^1 is greater than or smaller than β^2. A higher β, on the other hand, corresponds to a lower level of utility (see fn. 13). The expression (32), therefore, clearly demarcates the location of the optimal tax (or subsidy) between the two sectors: *the workers who are better off should be taxed, while the workers who are worse off should be subsidized. This result is independent of the behavioral responses in the economy, of the social valuation of investment, and of whether the peasants or the industrial workers are better off.* Note, however, that the *magnitude* of subsidy does depend on the social weight on investment.

The influence of the rural surplus response on the subsidy can be partly understood in the following manner. For the moment, assume that the social weights are constants. Then, a higher ε^1_{Qp} means that an increase in the price p leads to a higher loss (gain) to the government if p exceeds (is less than) η/δ. The government, therefore, would want to keep a lower absolute value of the rate of subsidy, as (32) indicates. This explanation, however, is only partial because our formulae for an optimum do not provide a closed-form solution for the rate of subsidy.

The formula for the optimal price is easily obtained by equating the right-hand sides of (32) and (33), and using (14):

$$(34)\quad px^2\left(\beta^1-\beta^2\right)=wL^2\left(\delta-\beta^2\right)\varepsilon_{wp}.$$

The above expression equates the net social gain from a price increase to the net social loss due to the price-induced wage increase. Rearranging the terms in (34), we obtain a remarkably simple rule to characterize the optimum.

---

[15] Note that the focus in this paper is on determining the properties of the optimal regime at any given point in time. An instructive and useful extension will be to find out how the optimum policy variables as well as the resulting welfare levels of individuals change over time. Such an analysis of the optimal time path is beyond the scope of the present paper.

[16] We use (4) and (7), and also the expressions for $\partial y^1/\partial p$, $\partial y^2/\partial p$, and $\partial y^2/\partial w$ from (2) and (5).

$$(35)\qquad \frac{(\delta-\beta^2)\varepsilon_{wp}}{(\beta^1-\beta^2)}=\alpha_x^2\equiv \text{budget share.}$$

Alternative Characterizations: In the above analysis we have employed the terms of trade as the instrument of control. This we have done to keep the analysis within the context of the scissors problem. There are several alternative ways, however, in which the present problem can be characterized.

First, consider the control of the nominal price of the urban good or the nominal urban wage, because it is through these variables that a terms-of-trade policy can actually be implemented. It can be verified that an increase in the nominal price of the urban good corresponds to a movement in the terms of trade against the rural sector, and that an increase in the nominal urban wage corresponds to a movement in the terms of trade in favor of the rural sector. Either one of these two nominal variables can therefore be used as the instrument of control, without affecting the results.

Second, the economic content of the analysis remains unchanged if either the level of investment, I, or the rural surplus, Q, is used as the independent policy instrument. In these cases the analysis will correspond to what is sometimes called the "investment problem," and the "marketed surplus" problem.

Third, our conclusions regarding the terms of trade remain essentially unaltered if the "modern" capital good produced in the urban sector can be productively employed in the rural sector, and if the government exercises the choice over the intersectoral allocation of investment. As a simple example, consider the case in which there is positive investment in both sectors and the rent on the rural capital accrues to the rural sector. Then it can be verified that pricing rules (32)–(35) continue to characterize the optimum. Similarly, our qualitative results on the terms of trade will not change if a time-dependent process of technical progress in production is incorporated in the model.

Finally, we can reinterpret our analysis in a decentralized economic setting. To see this, first imagine that the government instructs its public sector managers to maximize profits based on the nominal prices they face, but introduces either a commodity tax on the urban good or a wage tax. Naturally, then, the government can control the terms of trade at any level that it desires. In particular, the optimum which we analyzed earlier is implementable in this manner. Next, assume that the industrial production is privately owned, but that the government imposes a 100 percent profit tax, and that it also imposes one of the two taxes mentioned above. It follows that the desired public policy can be implemented through a private market equilibrium.[17]

VI. Unemployment and Labor Mobility in Developing Economies

Many developing economies are characterized by significant unemployment as well as by endogenous mobility of labor between rural and urban sectors. It is then necessary that the design of pricing policy in such economies should take into account its influence on labor mobility and unemployment.[18] In this section, therefore, we extend our basic model to include these additional aspects.

The population is now divided into three groups consisting of employed rural workers, employed urban workers, and unemployed workers. Their utilities are represented by $V^1(p,N^1)$, $V^2(p,w)$, and V^u, and their populations are denoted by N^1, N^2, and N^u, respectively. For simplicity, the present analysis abstracts from transfer arrangements from the employed to the unemployed workers, and assumes that the unemployed workers have zero income.

---

[17] The relationship between the optimal production and pricing in public enterprises and optimal taxation of private enterprises has been discussed in the standard tax literature. See, for instance, Stiglitz and Dasgupta.

[18] By the same token, the valuation of labor, i.e., the shadow wage of labor, will be influenced not only by the dependence of the two sectors through trade and resource transfers, but also by the pricing policies of the government. The existing literature on the shadow wage has not paid sufficient attention to these relationships. See our 1983a paper for a more complete analysis.

We posit that the mobility of labor across different groups of workers can be described through a reduced-form equation

$$(36) \qquad N^1 = N^1(p, w, N^2),$$

which gives the rural population as a function of the relative price, the urban wage, and urban employment. It follows then that the level of unemployment can be obtained as a function of the same variables, since

$$(37) \qquad N = N^1 + N^2 + N^u.$$

Note that our representation of labor mobility is quite general. Specific hypotheses concerning migration which have been proposed in the literature, for example, the Harris-Todaro hypothesis (John Harris and Michael Todaro, 1970), can be treated as special cases of the above.

Next, one needs to specify the determination of the urban wage. A number of alternative theories have been proposed in the literature (see Stiglitz, 1982a, b), and it is possible to examine the correct pricing policy for each of these theories of urban wage determination (see our 1983a paper for such an approach). For brevity, however, we follow here a simple hypothesis, according to which the urban wage is fixed, and the urban employment is controlled by the government. This representation of what the government can control in a developing economy conforms to the view that the urban wage is institutionally determined and also that much of the urban employment is in the public sector, over which the government exercises some control.

Note, however, that the two policy variables, p and N^2, cannot be controlled independently in the present problem. This is because a given combination of (p, N^2) determines a level of N^1 through (36), and the resulting set of variables cannot in general be expected to satisfy the quantity balance of the rural good, (10).

For later use, we define the following elasticities. Based on (36),

$$(38) \quad M_p = \frac{\partial \ln N^1}{\partial \ln p} \quad \text{and} \quad M_e = \frac{\partial \ln N^1}{\partial \ln N^2}$$

are the elasticities of the rural population with respect to the relative price and urban employment, respectively, and $\varepsilon^1_{Qa} = \partial \ln Q / \partial \ln a$ is the elasticity of rural surplus per rural worker with respect to land per rural worker. Further, $\hat{\varepsilon}^1_{Qp} = \partial \ln(N^1 Q)/\partial \ln p$ is elasticity of total rural surplus with respect to its price. It is simple to verify that

$$(39) \qquad \hat{\varepsilon}^1_{Qp} = \varepsilon^1_{Qp} + \left(1 - \varepsilon^1_{Qa}\right) M_p .$$

The relevant Lagrangian is defined by (31) in which

$$(40) \quad \psi = N^1 W(V^1) + N^2 W(V^2) + (N - N^1 - N^2) W(V^u).$$

The first-order conditions with respect to p and N^2 can be obtained easily. For brevity, we present and interpret only the first-order condition with respect to p, which after making substitutions which are similar to those encountered earlier, can be written as

$$(41) \quad s = \frac{(\beta^1 - \beta^2) + \phi M_p / pQ}{\delta\left[\varepsilon^1_{Qp} + \left(1 - \varepsilon^2_{Qa}\right) M_p + \varepsilon^2_{xp}\right]},$$

where

$$(42) \quad \phi = W(V^1) - W(V^u) - \beta^1 p X_a a.$$

Note immediately that (41) reduces to (32) if $M_p = 0$. This is precisely what we would expect. If the marginal migration is zero, then the rule for optimal pricing with endogenous migration is the same as the one in which populations are fixed.

The additional effects introduced by labor mobility can be understood intuitively by comparing (41) with (32). First note that the rural surplus response in the denominator of (41) is now redefined, according to (39), to include the effect of price on the size of rural population. Second, from (42), ϕ is the gain in the social welfare if one unemployed worker becomes employed in the rural sector. This includes the direct gain, $W(V^1) - W(V^u)$, and the indirect loss, $\beta^1 p X_a a$, due to the increased congestion on the rural land. The term containing ϕ in (41), therefore, represents the welfare gain from the price-induced labor mobility.

In the rest of this section, we provide a specific illustration of the above approach in which the migration is described by the Harris-Todaro hypothesis. According to this hypothesis, the unemployed workers are in the urban sector, and the probability that a migrant worker from the rural sector finds employment in the urban sector is $N^2/(N-N^1)$. The sectoral populations are determined by the equalization of the rural utility with the expected urban utility. That is, the expression (36) is specialized to

$$(N-N^1)V^1 = N^2V^2 + (N-N^1-N^2)V^u. \tag{43}$$

To ensure that N^u remains positive, it is explicitly assumed that $V^2 > V^1 > V^u$. This naturally restricts the range of prices within which the above hypothesis is meaningful. In addition, we make the following simplifying assumptions: (*i*) the rural land is not too scarce, that is, X_a and ε^1_{Qa} are negligible in magnitude, and (*ii*) the social welfare function is utilitarian, that is, $W(V) = V$, and $\beta^i = \lambda^i$. Because of the second assumption, our results do not depend on whether the social welfare function is defined over the *ex ante* or the *ex post* utilities of workers.

By perturbing (43) with respect to p and N^2, we find that

$$M_p > 0 \quad \text{and} \quad M_e < 0. \tag{44}$$

A result follows immediately. *An increase (decrease) in the price of agricultural output will be accompanied by an increase (decrease) in industrial employment.* This can be seen as follows. An increase in p increases the total rural surplus because $\hat{\varepsilon}^1_{Qp} > 0$, from (39) and (44). Also, an increase in p decreases the urban demand for the rural good. As a result, there will be an oversupply of the rural good. An increase in the urban employment will therefore be necessary to achieve a balance between the demand and supply of the rural good.

From the first-order conditions of optimality with respect to p and N^2, we obtain

$$s = \lambda^1 N/\delta N^1\left(\hat{\varepsilon}^1_{Qp} + \varepsilon^2_{xp}\right), \tag{45}$$

$$s = N^2L^2(w - Y_L)/N^1pQ(1 - M_e), \tag{46}$$

where $Y_L = \partial Y/\partial L^2$.

There are two results to be noted. First, $s > 0$ from (45). Therefore, *the optimal terms of trade will entail a subsidy to peasants and a tax on industrial workers.*

Second, *the optimal terms of trade will entail a level of industrial employment such that the industrial wage will exceed the marginal product of an industrial worker.* This follows from (44), (46), and $s > 0$. The idea behind this result is quite intuitive. The society in the present case is concerned only with the rural utility because this utility is the same as the expected utility of other workers. An increase in the rural price therefore raises the social welfare associated with consumption. From an earlier result, on the other hand, a higher rural price will be accompanied by an increase in the urban employment. The contribution of an additional urban worker to investment is $(Y_L - w)L^2$, which declines as urban employment increases because of the declining marginal product. The optimal terms of trade will thus involve a tradeoff between a gain in consumption and a loss in investment. The latter obviously implies that $Y_L < w$.

VII. Comments on the Soviet Debate

A significant shortcoming of the Soviet debate was the lack of attention given to the behavioral responses of peasants. This is especially surprising since the economic events faced by the early Soviet state (Dobb, ch. 7), as well as Vladimir Lenin's warnings,[19] had already pointed out the importance of incentives. Be that as it may, we have clearly demonstrated that the behavioral responses of peasants are central to any analysis of the scissors problem.

On the specific effects of turning the prices against peasants, we find that the government *can indeed* increase the rate of accumulation by imposing a price squeeze on peasants. This effect was originally claimed by Preobrazhensky, and his claim is valid, despite the doubts raised by subsequent researchers. But, the other important proposition made by Preobrazhensky, that the price squeeze can be operated without hurting the industrial proletariat, is invalid. A price squeeze on peasants will hurt the industrial proletariat, just as it will hurt the peasants.

---

[19] "...it is impossible to increase the production and collection of grain...except by improving the condition of the peasantry..." (1975b, p. 536).

It is of some interest to note here that much of the criticism faced by Preobrazhensky in the Soviet debate was that he was anti-peasant. "They accused him of favoring the 'exploitation' of the peasants, of advocating a kind of internal colonialism" (Alec Nove, 1965, p. *xi*). In fact, Preobrazhensky himself devoted much of his energies to trying to prove that he was not as anti-peasant as his book might have suggested at first (see his reply to Nikolai Bukharin and other opponents in the Appendix in his book). Some of this criticism might have been avoided, we suspect, had he not claimed (incorrectly, as demonstrated by our analysis) that the industrial proletariat would not have to pay any price for the state's accumulation.

Our analysis has also shown that the determination of the correct "tax" level or the correct size of the scissors depends on the social valuation of the welfare of the peasants and the industrial proletariat, as compared to the social valuation of investment (see equations (32)–(35)). On this score, it has often been believed that the peasants' welfare was irrelevant to the early Soviet state. Probably a better interpretation of the precollectivization debates might be that Preobrazhensky represented the lower end of the concern for the peasants, in contrast to Lenin who represented the middle position,[20] and to Bukharin who represented the higher end of the concern for the peasants.[21] On the other hand, it appears that the early Soviet leadership was fairly unanimous in placing a higher social weight on investment as compared to consumption, and in placing a higher social weight on the consumption of the industrial proletariat as compared to the consumption of the peasants.[22]

For this interpretation of the initial Soviet situation, our analysis shows that a movement in the terms of trade against peasants emerges as desirable, at least to that point where investment and the consumption of industrial protetariat have the same social weight (see equation (28)). The *direction* of change in the terms of trade, therefore, remains the same even though the relative concern for peasants might differ! The level of tax to be imposed on peasants, on the other hand, will differ according to the social concern for them: a higher concern will correspond to a lower tax (see equation (32)).

VIII. Extensions

There are two important aspects of a model of policy analysis such as the present one: the structure of the economy under consideration, and the instruments of policy which the government can (or cannot) employ. We have worked with a simple dual economy model, while ensuring an adequate representation of individuals' incentives. The terms of trade is the instrument of control on which we have focussed, although we have discussed a number of alternative ways in which the same control problem can be characterized. It is always possible to include *additional* instruments in an analysis, but one needs to exercise some care in doing so.

Take the example of the policy debates in the Soviet Union in the 1920's. A number of instruments of policy, other than the terms of trade, were brought up in these debates; among them were quantity controls, credit policy, foreign trade and borrowing, railway tariff, and printing money. It was widely acknowledged, however, that the opportunities provided by these additional instruments were at best quite limited, given the institutional limitations of the economy at that time. Further, it was also acknowledged that the inadequacies of the bureaucracy and the possibility of an underground economy were quite serious, and that these features should be considered in using any instrument which required a direct control of quantities.[23]

---

[20] See Lenin (1975a).

[21] Bukharin, in fact, exhorted the peasants to *enrich* themselves.

[22] Stated differently, it appears that there was unanimity on $\delta > \beta^2 > \beta^1$. But there were differences on the relative size of β^1 compared to β^2. It is obvious that, in the present context, the welfare evaluations are not based on an anonymous social welfare function.

[23] "Exchange is freedom to trade; it is capitalism. It is useful to us in as much as it will help us overcome the

The institutional constraints faced by many developing countries today are no less restrictive, probably much more so, than those faced by the early Soviet state. For such countries, therefore, it is desirable to focus on instruments like the terms of trade, which are less difficult to implement.

In developing economies with relatively greater institutional capabilities, however, many instruments of policy have been used which are relevant to the town vs. country considerations. Among the instruments which need to be examined are agricultural marketing boards and internal tax borders which exercise some regional control over prices and can tax the marketed surplus, and nonprice methods for the distribution of food in urban areas.[24,25]

IX. Conclusion

The problem of price scissors has remained a controversial issue ever since the Soviet leadership debated it in the 1920's. Also, it is a topic of substantial importance in much of today's developing world. Our analysis of the problem shows that some of the received wisdom on this issue is correct, while some of it is not correct. We conclude that the rate of accumulation in a socialist economy can indeed be increased by imposing a price squeeze on the rural sector. But, a price squeeze on the peasants leads to a decrease in the welfare of industrial workers, just as it leads to a decrease in the welfare of the peasants. We have also identified the critical parameters which influence the impact of the terms of trade on the state's surplus, and on the welfare of industrial workers. These parameters are the price response of the rural surplus, and the price and the income response of the urban consumption of the rural good.

Our analysis of desirable reforms in the terms of trade yields rules that are particularly simple to use, requiring only the knowledge of the social weights on the incomes of workers in the two sectors and of the social weight on investment. In many cases, the desirability of a shift in the terms of trade can be assessed simply by looking at the intertemporal tradeoffs (rather than by looking at the rural-urban tradeoffs). Finally, we have analyzed the optimal terms of trade. This analysis leads to a remarkably simple rule to determine which of the two sectors should be taxed and which should be subsidized.

We have then extended our analysis to incorporate unemployment and labor mobility, which are critical features of many of today's developing economies. Using a simple model with rigid industrial wage, we show that the optimal terms of trade will entail a subsidy to the rural sector and a tax on the urban sector. Further, the corresponding level of urban employment will be such that the urban wage will exceed the marginal product of an urban worker.

We also discuss a number of other ways in which the present approach can be expanded to suit the varying economic structures and institutional capabilities of different economies. These extensions will naturally modify the precise form of the results obtained here. We believe, however, that the central economic effects which we have brought out in our analysis, focussing on the consequences of pricing policy (terms of trade) on incentives in the rural and the urban sectors, and the general equilibrium effects of a price change on the welfare of individuals and on accumulation, will remain valid.

REFERENCES

Atkinson, Anthony B. and Stiglitz, Joseph E., *Lectures on Public Economics*, New York: McGraw-Hill, 1980.

Barnum, Howard N. and Squire, Lyn, *A Model*

---

dispersal of the small producer, and to a certain degree combat the evils of bureaucracy" (Lenin, 1975b, p. 555).

[24] Another issue which arises repeatedly in policy discussions is the role and the nature of correct pricing and taxation policies in those open developing economies which not only have the special features associated with dualism, but which in addition face restrictions on the external trade of some of the goods they produce, as well as on the external borrowing which is available to them. For an analysis of many of these issues, see our 1983b paper.

[25] See Sah (1982) for a comparison of the performance of several different nonprice instruments to allocate a good within a single sector with many consumers.

of an Agricultural Household, Baltimore: Johns Hopkins University Press, 1979.

Behrman, Jere R., *Supply Response in Underdeveloped Agriculture: A Case Study of Four Major Annual Crops in Thailand, 1937–1963*, Amsterdam: North-Holland, 1968.

Bukharin, Nikolai I., *Economics of the Transformation Period*, New York: Bergman Publishers, 1971.

Dixit, Avinash K., "Marketable Surplus and Dual Development," *Journal of Economic Theory*, August 1969, *1*, 203–19.

______, "Models of Dual Economies," in James A. Mirrlees and Nicholas H. Stern, eds., *Models of Economic Growth*, New York: Wiley & Sons, 1973, 325–52.

Dobb, Maurice, *Soviet Economic Development since 1917*, London: Routledge and Kegan Paul, 1966.

Erlich, Alexander, *The Soviet Industrialization Debate*, Cambridge: Harvard University Press, 1960.

Harris, John and Todaro, Michael, "Migration, Unemployment, and Development," *American Economic Review*, March 1970, *60*, 126–42.

Hornby, J. M., "Investment and Trade Policy in the Dual Economy," *Economic Journal*, March 1968, *78*, 96–107.

Lau, Lawrence J., Lin, Wuu-Long and Yotopoulos, Pan A., "The Linear Logarithmic Expenditure System: An Application to Consumption-Leisure Choice," *Econometrica*, July 1978, *46*, 843–68.

Lenin, Vladimir, (1975a) "Economics and Politics in the Era of the Dictatorship of the Proletariat," in *Selected Works*, Vol. 3, Moscow: Progress Publishers, 1975, 230–38.

______, (1975b) "The Tax in Kind," in *Selected Works*, Vol. 3, Moscow: Progress Publishers, 1975, 526–56.

Lipton, Michael, *Why Poor People Stay Poor: Urban Bias in World Development*, Cambridge: Harvard University Press, 1977.

Marx, Karl, *Capital*, Vol. 1, New York: Vintage Books, 1967.

Mitra, Ashok, *Terms of Trade and Class Relations*, London: Frank Cass, 1977.

Nove, Alec, "Introduction," in Evgeny Preobrazhensky, ed., *The New Economics*, Oxford: Clarendon Press, 1965.

Peterson, Willis L., "International Farm Prices and the Social Cost of Cheap Food Policies," *American Journal of Agricultural Economics*, February 1979, *61*, 12–21.

Preobrazhensky, Evgeny, *The New Economics*, Oxford: Clarendon Press, 1965.

Sah, Raaj Kumar, "Performance of Coupons, Queues, and Rations in Managing Scarcities," CADE Discussion Paper 82-48, Department of Economics, University of Pennsylvania, 1982.

______ **and Stiglitz, Joseph E.,** (1983a) "The Social Cost of Labor, and Project Evaluation: A General Approach," mimeo., Department of Economics, Princeton University, 1983.

______ **and** ______, (1983b) "The Town, the Country, and the Rest of World: Economics of Public Marketing Boards in Open Developing Economies," mimeo., Department of Economics, Princeton University, 1983.

Schultz, Theodore W., *Distortions of Agricultural Incentives*, Bloomington: Indiana University Press, 1978.

Spulber, Nicholas, *Foundations of Soviet Strategy for Economic Growth*, Bloomington: Indiana University Press, 1964.

Stalin, Joseph, *Problems of Leninism*, Moscow: Progress Publishers, 1954.

Stiglitz, Joseph E., (1982a) "The Structure of Labor Markets and Shadow Prices in LDCs," in Richard H. Sabot, ed., *Migration and the Labor Market in Developing Countries*, Boulder: Westview Press, 1982.

______, (1982b) "Alternative Theories of Wage Determination and Unemployment: the Efficiency Wage Model," in Mark Gersovitz et al., eds., *The Theory and Experience of Economic Development*, London: George Allen and Unwin, 1982.

______ **and Dasgupta, Partha S.,** "Differential Taxation, Public Goods, and Economic Efficiency," *Review of Economic Studies*, April 1971, *38*, 151–74.

Trotsky, Leon, *1905*, New York: Random House, 1971.

Yotopoulos, Pan A. and Nugent, Jeffrey B., *Economics of Development*, New York: Harper and Row, 1976.

Marx and Malthusianism: Marx's Secular Path of Wages

By SAMUEL HOLLANDER*

Karl Marx identified the Malthusian population doctrine with the "iron law of wages" and read it as precluding improvement under any form of institutional arrangement (for Marx's position, see William Baumol, 1983, p. 306). He himself insisted that "every special historic mode of production has its own special laws of population, historically valid within its limits alone. An abstract law of population exists for plants and animals only, and only insofar as man has not interfered with them" (*Capital*, I, 1954, p. 632). Marx does not seem to have done justice to Malthus' *Essay on Population* which, far from maintaining the inevitability of excessive population growth, insisted upon a reduced birth rate as a means of *preventing* falling wages—a policy particularly necessary since Malthus also championed steps to reduce the mortality rate. (Textual justification for these assertions regarding Malthus will be found in David Levy, 1978, and my forthcoming articles.) Poverty was no more inevitable under a communist regime although Malthus feared that the required population control would entail severe constraints on personal freedom. But whether or not Marx was justified in his description of the *Essay* as a "libel on the human race" (Saul Padover, 1979, p. 192), it can be shown that his own celebrated "immiseration" doctrine involving the Reserve Army of Unemployed turns strategically upon the demographic variable. In this paper, I bring to light a feature of the Marxian growth model that has received little attention in the literature.[1]

Section I treats the relation between the orthodox "subsistence" wage—that wage at which population growth ceases—and Marx's "value of labour power." The potential for population growth, it transpires, is built into the latter variable, the subsistence wage (physiologically determined for Marx) constituting its lower limit. I then document (Section II) Marx's adherence to a secular path of wages which tends towards the subsistence level defined as above in orthodox fashion. The underlying cause of the decline is shown to be a continually decelerating

---

*Department of Economics, University of Toronto, 150 St. George Street, Toronto, Ontario, Canada M5S 1A1. For helpful if sometimes severe criticism, my warm thanks to W. J. Baumol, M. Bronfenbrenner, E. Forget, T. Kompas, D. Levy, M. C. Marcuzzo, and A. Rosselli.

[1] My general position has much in common with that of G. S. L. Tucker (1961, pp. 265, 267–68). Fred Gottheil (1966, p. 155) recognizes, but minimizes the significance of, the role of population growth as a cause of excess labor supply in Marx's analysis. Wally Seccombe remarks correctly that while Marx rejected any "natural or eternal law of human population growth or overpopulation," he "did not consistently uphold the principle of historical specificity with regard to population patterns in *Capital*, but often slipped back into the naturalist discourse which he criticized Malthus for" (1983, p. 32). Paul Sweezy categorically asserts that though "Marx never wrote much about the factors which determine the size of the population... this much is certain, that he had no use whatever for the Malthusian theory or any of its variants" (1956, p. 86). Moreover, Sweezy maintains that "the principle of the reserve army is independent of any particular population assumption, that it works equally well with a stationary or even a declining population. In this we have one of the decisive differences between Marx and his predecessors in the classical school..." (p. 89); "whatever presumption we care to make about factors outside the system (population), the fact remains that it contains within itself a mechanism for regulating the wage level and hence for maintaining profits" (p. 91). Yet rather surprisingly Sweezy does not deny "the practical and theoretical significance [for Marx] of the rate of population growth. The problem acquires great importance on a somewhat lower level of abstraction" (p. 222). David Harvey writes of "Marx's rather surprising failure to undertake any systematic study of the processes governing the production and reproduction of labour power itself.... This omission is, perhaps, one of the most serious of all the gaps in Marx' own theory, and one that is proving extremely difficult to plug..." (1982, p. 163). But, like Sweezy, he also maintains that although "the sociological, demographic and geographical aspects of labour supply are important for any general theory of accumulation," they can be set aside as a first approximation considering Marx's primary objective to show "that if misery, poverty and unemployment are found under capitalism, then they have to be interpreted as the product of this mode of production and not attributed to 'nature'" (p. 166).

growth rate of labor demand (reflecting the conversion of "variable" into "constant" capital)—but not an absolute fall—in the face of ongoing, possibly accelerating, population expansion.[2] Some attention is paid to the place of the Industrial Reserve Army in the analysis (Section III), and I conclude with reflections on Marx's neglect of population control as a measure in the prevention of falling real wages.

I. The Subsistence Wage and the Value of Labor Power

To clarify the Marxian growth path of wages, the relation between the orthodox subsistence wage and Marx's value of labor power must be first carefully defined.[3] In his speech on "Wages, Price and Profit" of 1865, market wages are said to tend "in the long run" towards equality with "the value of labour power." This latter comprises a "physical element"—corresponding to "the necessaries absolutely indispensable for living and multiplying"—which forms the minimum limit to the wage and is represented as an (apparently unchangeable) physiological quantum; and a cultural element reflecting the "traditional standard of life," which is represented as variable:

> ...as with all other commodities, so with labour, its *market price* will, in the long run, adapt itself to its *value*; that, therefore, despite all the ups and downs, and do what he may, the working man will, on an average, only receive the value of his labour, which resolves into the value of his labouring power, which is determined by the value of the necessaries required for its maintenance and reproduction, which value of necessaries finally is regulated by the quantity of labour wanted to produce them.
>
> But there are some peculiar features which distinguish the *value of labouring power, or the value of labour*, from the values of all other commodities. The value of the labouring power is formed by two elements—the one merely physical, the other historical or social. Its *ultimate limit* is determined by the *physical* element, that is to say, to maintain and reproduce itself, to perpetuate its physical existence, the working class must receive the necessaries absolutely indispensable for living and multiplying. The *value* of these indispensable necessaries forms, therefore, the ultimate limit of the *value of labour*....
>
> Besides this mere physical element, the value of labour is in every country determined by a *traditional standard of life*. It is not mere physical life, but it is the satisfaction of certain wants springing from the social conditions in which people are placed and reared up. The English standard of life may be reduced to the Irish standard; the standard of life of a German peasant to that of a Livonian peasant....
>
> This historical or social element, entering into the value of labour, may be expanded, or contracted, or altogether extinguished, so that nothing remains but the *physical limit*. [pp. 441–42]

The "physical element" thus permits labor merely "to perpetuate its physical existence"; at the physical limit the wage is just equal to an amount "necessary for the physical perpetuation of the race" (p. 443). Population (labor supply) at that wage is merely replaced but not expanded. The "ultimate limit" to the wage corresponds therefore to the orthodox subsistence wage—defined as that wage assuring zero population growth. Now in this same context (p. 445), we find the presumption of on-going contemporary population growth—Marx refers to the "English capital having grown in the last twenty years so much quicker than English population"—although (long-run) wages are assumed equal to the value of labor power. And this is the essential point: for Marx a wage equal to the value of labor power is consistent with net population growth only when allowance is made therein for the cultural element. (And here is a source of confusion, since for the orthodox economists the subsistence, or zero-growth, wage itself contained a cultural element.)

---

[2] My discussion is limited to the advanced industrial sector. Agriculture is treated differently, for there, Marx maintained, particularly extensive capital conversion had entailed *absolute* decline in the demand for labor, and the wage rate is represented to be already at subsistence (*Capital*, I, pp. 640f). Similarly, "domestic industry supernumeraries"—those displaced from the handicraft and decaying manufacturing sectors, as well as from agriculture—are said to be earning minimum wages. It is, in short, the downward wage trend within the advanced industrial sector that is my concern.

[3] References to the value of labor power throughout this paper are to commodity wages. Changes therein due to altered costs of producing wage goods are not our concern.

Let us turn for elucidation of this key issue to *Capital*. There again the definition of the minimum wage will be found with reference to constancy of population:

> The labour-power withdrawn from the market by wear and tear and death, must be continually replaced by, at the very least, an equal amount of fresh labour-power. Hence the sum of the means of subsistence necessary for the production of labour-power must include the means necessary for the labourer's substitutes, *i.e.*, his children, in order that this race of peculiar commodity-owners may perpetuate its appearance in the market.
> [Vol. I, p. 172; cf. p. 572]

Now, in this particular context, Marx stressed the constancy of the value of labor power at any particular state of historical development, by which expression he intended (consistently with the speech) not merely the minimum wage but the minimum supplemented by the cultural element—the laborer's "necessary" in addition to his "natural" wants. Here, too, he emphasized that the cultural element is governed by the habits under which "the class of free labourers has been formed"—contrasting with the natural element which varies "according to the climatic and other physical conditions of the country; confirming its physiological character" (*Capital*, I, p. 171; cf. *Marx-Engels*, p. 442). Similarly: "The value of labour-power is determined by the value of the necessaries of life habitually required by the average labourer. The quantity of these necessaries is known at any given epoch of a given society, and can therefore be treated as a constant magnitude" (*Capital*, I, p. 519). The same theme is restated in Volume III in a passage which even refers to the physical minimum as a "natural law":

> Wages...are regulated on the one hand by a natural law; their lower limit is determined by the physical minimum of means of subsistence required by the labourer for the conservation of his labour-power and for its reproduction; i.e., by a definite quantity of commodities.... The actual value of his labour power deviates from this physical minimum; it differs according to climate and level of social development; it depends not merely upon the physical, but also upon the historically developed social needs, which become second nature. But in every country, at a given time, this regulating average wage is a given magnitude. [p. 837]

In the full analysis of the growth process, the value of labor power is treated (as in the speech of 1865) as a variable—this is, in fact, what the falling wage trend so much emphasized in *Capital* entails. At this point, it is the Marxian position that a wage equal to the supposedly known value of labor power is consistent with population growth that I wish to establish. On this Marx is quite clear, both prior to the analysis of a changing organic composition of capital in Volume I, and in Volume III. The wage-earning class under capitalism earns "ordinary wages" which "suffice, not only for its maintenance, but for its increase":

> To accumulate it is necessary to convert a portion of the surplus-product into capital. But we cannot...convert into capital anything but such articles as can be employed in the labour-process (*i.e.*, means of production), and such further articles as are suitable for the sustenance of the labourers (*i.e.*, means of subsistence)....
>
> Now in order to allow of these elements actually functioning as capital, the capitalist class requires additional labour. If the exploitation of the labourers already employed do not increase, either extensively or intensively, then additional labour-power must be found. For this the mechanism of capitalist production provides before-

> hand, by converting the working-class into a class dependent on wages, a class whose ordinary wages suffice, not only for its maintenance, but for its increase. It is only necessary for capital to incorporate this additional labour-power, annually supplied by the working-class in the shape of labourers of all ages, with the surplus means of production comprised in the annual produce, and the conversion of surplus-value into capital is complete.
> [*Capital*, I, pp. 580–81]

> The working-class must find at least the same quantity of necessities at hand if it is to continue living in its accustomed average way, although they may be more or less differently distributed among the different kinds of commodities. Moreover, there must be an additional quantity to allow for the annual increase of population.
> [*Capital*, III, pp. 184–85]

In a growing system, then, the long-run wage rate, while equal to the value of labor power, exceeds the (physiological) subsistence rate at which population growth ceases; *population growth is, as it were, built into the value of labor power.*

It is in these terms that an otherwise severe inconsistency can be resolved, namely assertions to the effect that both wage increases and wage decreases are consistent with an increased rate of population growth. For in a discussion of "absolute over-production of capital" (upward pressure on wages exerted by particularly rapid accumulation), reference is made to a consequential acceleration of the population growth rate:

> Prosperity would have led to more marriages among labourers and reduced the decimation of offspring. While implying a real increase in population, this does not signify an increase in the actual working population. But it affects the relations of the labourer to capital in the same way as an increase in the number of actually working labourers would have affected them.
> [*Capital*, III, pp. 249–50]

Yet we also read that a fall in wages—due to the capital conversion process, a matter I postpone for the moment—would "be a breeding ground for a really swift propagation of the population, since under capitalist production misery produces population":

> [The] growth of the means of production includes the growth of the working population, the creation of a working population, which corresponds to the surplus-capital, or even exceeds its general requirements, thus leading to an over-population of workers. A momentary excess of surplus-capital over the working population it has commandeered, would have a two-fold effect. It would, on the one hand, by raising wages, mitigate the adverse conditions which decimate the offspring of the labourers and would make marriages easier among them, so as gradually to increase the population. On the other hand, by applying methods which yield relative surplus-value (introduction and improvement of machinery) it would produce a far more rapid, artificial relative over-population, which in its turn, would be a breeding ground for a really swift propagation of the population, since under capitalist production misery produces population.
> [p. 214]

I suggest the following possible solution to the apparent paradox consistently with the earlier exegesis: To each given "value of labour power" or "standard of life," there corresponds a specific growth rate of population; a wage increase, given the standard, will stimulate an increased population growth rate, and a wage decrease a decline. This much for fluctuations of the wage about a given standard. But should the fall in the wage entail a fall in the standard itself, then matters are very different; at the lower standard the population growth rate may remain constant, or, indeed, may even rise compared to the original level if the degradation in standards is sufficiently marked.[4]

---

[4] It is possible that the reference to misery as "a breeding ground" for rapid population growth applied specifically to domestic industry. For in the first volume we encounter a similar statement in that particular context:

II. The Secular Path of Wages

That wages for Marx can oscillate about the value of labor power is clear (*Capital*, III, p. 201). But that is not the concern here. Setting aside wage fluctuations, the wage reflects the value of labor power. What, now, of its secular course?

In the earliest stages of capitalism, accumulation proceeds with organic composition unchanged, so that the rate of labor demand is proportionate to the rate of accumulation (*Capital*, I, p. 621). If "the requirements of accumulating capital... exceed the increase of labour-power or of the number of labourers; the wage rate will rise" (p. 613; cf. p. 618). *It is thus to deviations between the growth rates of labor demand* (*accumulation*) *and supply that Marx attributes the wage movement.* Any upward movement in the wage is, however, held in check by the dampening effects on the rate of accumulation exercised by the falling return on capital:

> But as soon as this diminution touches the point at which the surplus-labour that nourishes capital is no longer supplied in normal quantity, a reaction sets in: a smaller part of revenue is capitalized, accumulation lags, and the movement of rise in wages receives a check. The rise of wages therefore is confined within limits that not only leave intact the foundations of the capitalistic system, but also secure its reproduction on a progressive scale.
> [*Capital*, I, p. 620]

There is no talk here of excess labor supply; merely a limit to the extent the wage can rise.

I now introduce the complexity of a changing organic composition of capital, Marx's main preoccupation—"that change in the technical composition of capital by which the variable constituent becomes always smaller and smaller as compared with the constant" (p. 624). Exogenous technical change takes precedence by far over wage-induced substitution against labor, although Marx is not always crystal clear regarding his meaning; frequently citing Barton and Ricardo who in fact dealt with both exogenous and wage-induced adoption of machinery (see for example, *Capital*, I, pp. 430n, 432n, 438n, 631n; *Marx-Engels*, I, p. 445). (On Barton and Ricardo, see my 1971 article.) But the weight of emphasis throughout is on technological change introduced by individual capitalists to raise productivity and cut costs in a scramble for markets—attempts sooner or later adopted by the industry as a whole—quite apart from any preceding wage increase: "The battle of competition is fought by cheapening of commodities. The cheapness of commodities depends, *caeteris paribus*, on the productiveness of labour, and this again on the scale of production" (*Capital*, I, p. 626; cf. pp. 430–31 and 449f). And so machine-intensive technologies may be introduced even when wages have fallen to drastically low levels (p. 470). The central argument of the section treating the Progressive Production of Relative Surplus-Population or Industrial Reserve Army (pp. 628f), based on the "law" of the progressive increase in constant relative to variable capital, proceeds indeed on the explicit assumption of constant wages—"that the increase or diminution of the variable capital corresponds rigidly with the increase or diminution of the number of labourers employed" (p. 635).[5]

With this established, note that it is not merely the additions to capital which embody the new technologies that entail altered composition; the entire capital stock ultimately comes to be transformed (p. 628). But

---

> ... not only the number of births and deaths, but the absolute size of the families stand in inverse proportion to the height of wages, and therefore to the amount of means of subsistence of which the different categories of labourers dispose. This law of capitalist society would sound absurd to savages, or even civilized colonists. It calls to mind the boundless reproduction of animals individually weak and constantly hunted down.
> [*Capital*, I, p. 643]

In this case, the wage must have fallen to very low levels indeed before further reductions generate population increase.

[5] It is an essential part of the interpretation by Sweezy that the capital conversion process (and the generation of the reserve army of unemployed) reflects substitution against labor in response to a preceding wage increase (pp. 87–88). The only citation given in support is a passage relating to the response by employers to a sudden and unusual fall in English agriculture labor, 1849–59 (see Section IV below).

even so an absolute fall in the demand for labor in consequence of the altered composition is not what Marx had in mind. The net outcome of an expanding aggregate capital stock and altering capital composition is said to be expanded demand for labor, though at a decelerating rate:

> Since the demand for labour is determined not by the amount of capital as a whole, but by its variable constituent alone, that demand falls progressively with the increase of total capital, instead of, as previously assumed, rising in proportion to it. It falls relatively to the magnitude of the total capital, and at an accelerated rate, as this magnitude increases. With the growth of the total capital, its variable constituent or the labour incorporated in it, also does increase, but in a constantly diminishing proportion. The intermediate pauses are shortened, in which accumulation works as simple extension of production, on a given technical basis. It is not merely that an accelerated accumulation of total capital, accelerated in a constantly growing progression, is needed to absorb an additional number of labourers, or even, on account of the constant metamorphosis of old capital, to keep employed those already functioning. In its turn, this increasing accumulation and centralization becomes a source of new changes in the composition of capital, of a more accelerated diminution of its variable, as compared with its constant constituent.
>
> [pp. 629–30; cf. pp. 623, 642]

The picture is one of accelerating accumulation and a continually decreasing rate of increase of labor demand, indeed an accelerating deceleration (p. 631). But that Marx presumed some net expansion of labor demand is certain.[6] The point is frequently reiterated. "In the centres of modern industry—factories, manufactures, ironworks, mines, etc.—the labourers are sometimes repelled, sometimes attracted again in greater masses, the number of those employed increasing on the whole, although in a constantly decreasing proportion to the scale of production" (p. 641). And in Volume III: "...it is but a requirement of the capitalist mode of production that the number of wage-workers should increase absolutely, in spite of its relative decrease." Indeed, were it otherwise there would be a revolution—one of the most remarkable pronouncements in *Capital*: "A development of productive forces which would diminish the absolute number of labourers, i.e., enable the entire nation to accomplish its total production in a shorter time span, would cause a revolution, because it would put the bulk of the population out of the running" (*Capital*, III, p. 258; also *Marx-Engels*, I, p. 445).

Labor demand, therefore, is *increasing*. What then is the problem? Why the downward pressure on wages? The answer evidently lies in the fact of on-going population growth discussed above—in short, the "additional number of labourers" just cited from *Capital*, I (p. 630), or "the natural increase" of population alluded to in the same context (p. 635). For "what are set free are not only labourers immediately turned out by the machines, but also their future substitutes in the rising generation, and the additional contingent, that with the usual extension of trade on the old basis would be regularly absorbed" (p. 639).[7] *And the consequence of the deviation between the growth rates of labor demand and supply is a reduction in real wages*:

> The first word of this adaptation [of variable into constant capital] is the creation of a relative surplus-population, or industrial reserve army. Its last word is the misery of constantly extending strata of the active army of labour, and the dead weight of pauperism.... [The] higher the productiveness of labour, the greater is the pressure of the labourers on the means of employment, the more precarious, therefore, becomes their condition of existence, viz., the sale of their own labour-power

---

[6] Though not necessarily in individual industries (see *Capital*, I, p. 448).

[7] On the presumed increase in population, see also *Capital* (I, pp. 441, 454; III, p. 799); *Grundrisse* (pp. 608, 764, 771).

> for the increasing of another's wealth, or for the self-expansion of capital. [pp. 644–5][8]

This picture should be contrasted with that in the colonies (pp. 769–70; also *Marx-Engels*, I, p. 444); or with early capitalism, with their upward wage trends due to deviations in factor growth rates favoring labor.

The problem is summarized in a passage which appears at first sight rather mysterious since reference is made to "an *apparently* absolute increase of the labouring population" when we know that Marx has presumed throughout an *actual* absolute increase in population (even in employment, considering the net expansion of labor demand):

> [The] accelerated relative diminution of the variable constituent, that goes along with the accelerated increase of the total capital, and moves more rapidly than this increase, takes the inverse form, at the other pole, of an apparently absolute increase of the labouring population, an increase always moving more rapidly than that of the variable capital or the means of employment. But in fact, it is capitalistic accumulation itself that constantly produces, and produces in the direct ratio of its own energy and extent, a relatively redundant population of labourers, *i.e.*, a population of greater extent than suffices for the average needs of the self-expansion of capital, and therefore a surplus-population. [p. 630; cf. pp. 633, 645]

The mystery is resolved if we understand Marx's intention to be the denial that the problem of excess labor supply was "due to" population growth as the disturbing cause. That on-going population growth occurs is not in doubt. But the problem of excess labor supply has its immediate source in a falling rate of growth of labor demand.

Marx's apparent denial of an absolute increase in population emerges also during the course of a formal critique of Malthus, the complaint that while Malthus had recognized "over-population as a necessity of modern industry," he had "after his narrow fashion" explained it "by the absolute over-growth of the labouring population, not by their becoming relatively supernumerary" (p. 634). But Marx's position here may be appreciated as a denial of any problem of absolute population growth *relative to land* and manifested in diminishing agricultural productivity. This is precisely how the matter is summarized in *Capital*, Volume III, in passages unambiguously allowing for on-going population expansion:

> ...the possibility of a relative surplus of labouring people develops proportionately to the advances made by capitalist production not because the productiveness of social labour *decreases*, but because it *increases*. It does not therefore arise out of an absolute disproportion between labour and the means of subsistence, or the means for the production of these means of subsistence, but out of disproportion occasioned by capitalist exploitation of labour, a disproportion between the progressive growth of capital and its relatively shrinking need for an increasing population. [pp. 217–18]

> ...as the capitalist mode of production develops, an ever larger quantity of capital is required to employ the same, let alone an increased, amount of labour-power. Thus, on a capitalist foundation, the increasing productiveness of labour necessarily and permanently creates a seeming over-population of labouring people. [p. 219]

The fact of on-going population growth is thus beyond doubt, despite appearances on occasion, as is the relatively slower growth of labor demand. And, to repeat (in terms of

---

[8] Roman Rosdolsky, who ignores in this context the issue of population pressure, claims that Marx refers here only to the growing misery of the "lazarus layers of the working class," *not* the general body of labor (1977, pp. 302–03). But Marx's reference to "the misery of constantly extending strata of the active army of labour..." and the growing "precarious[ness]" of "their condition of existence" (a reference to "labourers" without qualification), suggests otherwise. (Rosdolsky omits these statements from n. 57.)

Marx's speech of 1865), the consequence is a reduction in the value of labor power "to its minimum limit":

> These few hints will suffice to show that the very development of modern industry must progressively turn the scale in favour of the capitalist against the working man, and that consequently the general tendency of capitalist production is not to raise, but to sink the average standard of wages, or to push the *value of labour* more or less to its *minimum limit*.
> [*Marx-Engels*, I, p. 446]

Rosdolsky (p. 303, n56) has claimed that the foregoing passage must be understood subject to the absence of trade union counterpressure. That this is so, runs the argument, is implied by the continuation of the passage:

> Such being the tendency of *things* in this system, is this saying that the working class ought to renounce their resistence against the encroachments of capital, and abandon their attempts at making the best of the occasional chances for their temporary improvement? If they did, they would be degraded to one level mass of broken wretches past salvation. [p. 446]

Now it is true enough that the wage rate is the outcome of a "continuous struggle between capital and labour, the capitalist constantly tending to reduce wages to their physical minimum, and to extend the working day to its physical maximum, while the working man constantly presses in the opposite direction" so that "the matter resolves itself into a question of the respective powers of the combatants" (p. 443). But with which party does the advantage lie? Marx is perfectly explicit that trade unions can only check but cannot prevent the downward trend:

> ...the working class ought not to exaggerate to themselves the ultimate working of these everyday struggles. They ought not to forget that they are fighting with effects, but not with the causes of those effects; that *they are retarding the downward movement but not changing its direction*; that they are applying palliatives, not curing the malady. [p. 446; emphasis added]

Some confusion has also been created in the secondary literature by Marx's famous statement that "in proportion as capital accumulates, the lot of the labourer, be his payment high or low, must grow worse" (*Capital*, I, p. 645). It is sometimes said that this statement allows one to conceive of increasing "immiseration" despite rising real wages.[9] But this interpretation does not allow for the extensive textual evidence given above pointing to declining real earnings in the ordinary sense of that term. It is more likely that Marx intended a decline in the general wage rate which affects all classes of laborers—those high on the wage scale as well as those at a lower level. (On sectoral differentials, cf. pp. 653, 659, 667f.)

III. The Industrial Reserve Army and Cyclical Wage Fluctuations

The falling secular wage does not act as a corrective to excess labor supply insofar as the population growth rate is concerned—on the contrary, the fall in standards may be such that the rate of growth of population not only remains unchanged at lower wages but even accelerates. On the other hand, workers displaced are to some extent reabsorbed at lower wages into labor-intensive sectors (see *Capital*, III, p. 232). Yet any stimulation to labor demand could at best act as a partial corrective. For the conversion of variable into constant capital is an ongoing process, proceeding at an ever-faster rate which, always bearing in mind steady and possibly accelerating population growth, aggravates the excess labor supply and assures continuous downward pressure on the wage; the problem is not defined in the static terms appropriate to a once-and-for-all disturbance.

---

[9] For example, Rosdolsky states that if by the "accumulation of misery" (*Capital*, I, p. 645), Marx referred to the working class as a whole, "one would have to suppose that [he] expected this 'ignorant, brutalized and morally degraded' working class [to which he referred in our present context] to establish socialism—something which might perhaps be asserted by Bakunin, but not by Marx!" (pp. 303–04). But I have shown that downward trend in wages to be a quite general phenomenon for Marx. Rosdolsky avoids my conclusion, in part, by maintaining that any downward pressure on wages may be passed on to underdeveloped areas of the world (pp. 307f). See also Ernest Mandel: the tendency of average wages to fall "applies only to capitalist society taken as a whole, that is, on the *world* scale..." the burden being transferred "to the countries of the 'Third World'" (1971, p. 149).

But while the trend path of wages is certainly downward, allowance must be made for cyclical fluctuations about the trend, and the possibility of periods of constant wages. These short-term plateaus must not be mistaken for the secular trend itself, as it may be easy to do (for example, Michio Morishima, 1973, pp. 129f). I touch here on the role of the Reserve Army of Unemployed in providing a source of available labor to satisfy cyclical bursts of activity:

> But if a surplus labouring population is a necessary product of accumulation or of the development of wealth on a capitalist basis, this surplus-population becomes, conversely, the lever of capitalistic accumulation, nay, a condition of existence of the capitalist mode of production. It forms a disposable industrial reserve army, that belongs to capital quite as absolutely as if the latter had bred it at its own cost. Independently of the limits of the actual increase of population, it creates, for the changing needs of the self-expansion of capital, a mass of human material always ready for exploitation. With accumulation, and the development of the productiveness of labour that accompanies it, the power of sudden expansion of capital grows also; ...there must be the possibility of throwing great masses of men suddenly on the decisive points without injury to the scale of production in other spheres. Over-population supplies these means. The course characteristic of modern industry, viz. a decennial cycle (interrupted by smaller oscillations), of periods of average activity, production of high pressure, crisis and stagnation, depends on the constant formation, the greater or less absorption, and the reformation of the industrial reserve army or surplus-population.
>
> [*Capital*, I, pp. 632–33; cf. III, p. 258]

This pattern stands in sharp contrast to early capitalist experience where net accumulation, and the corresponding net increase in labor demand, although of slow growth, "found a check in the natural limits of the exploitable labouring population..." (p. 633). What was now occurring was "expansion by fits and starts of the scale of production" calling for "an increase in the number of labourers independently of the absolute growth of the population."

The argument to this point implies that cyclical variations in employment can occur without pressure on wages. Yet this requires qualification. Wages may yet fluctuate (within limits) governed by the extent of the excess labor supply:

> Taking them as a whole, the general movements of wages are exclusively regulated by the expansion and contraction of the industrial reserve army, and these again correspond to the periodic changes of the industrial cycle. They are, therefore, not determined by the variations in the absolute number of the working population, but by the varying proportions in which the working-class is divided into active and reserve army, by the increase or diminution in the relative amount of the surplus-population, by the extent to which it is now absorbed, now set free. For Modern Industry with its decennial cycles and periodic phases, which, moreover, as accumulation advances, are complicated by irregular oscillations following each other more and more quickly, that would indeed be a beautiful law, which pretends to make the action of capital dependent on the absolute variation of the population, instead of regulating the demand and supply of labour by the alternate expansion and contraction of capital, the labour-market now appearing relatively under-full, because capital is expanding, now again over-full, because it is contracting.
>
> [*Capital*, I, p. 637; cf. pp. 639, 769]

Conceivably, the wage rate may rise in consequence of cyclical bursts of activity if the pressure is sharp enough (*Capital*, III, p. 246). (Wages may indeed rise despite the existence of some unemployment where the surplus labor is of low efficiency; see p. 251; cf. p. 240.) I shall return to the passage from Volume I in the next section.

IV. Marx's Objections to Malthus Considered

For both Malthus and Marx, the growth of labor demand tends to decline secularly though for very different reasons since, for Marx, land scarcity—diminishing returns—is not an issue. My concern is Marx's presumption, in contrast to Malthus, of a *necessarily* excessive growth rate of labor supply forcing wages downwards. His purpose, unlike that of Malthus (or Ricardo), was not, of course, the formulation of a reform program designed to improve the living standards of the masses within capitalism. Nonetheless, since his forecast of falling real wages hinged upon excessive population pressure, we are obliged to look more closely into his neglect of the "Malthusian" solution. Here two issues must be distinguished: first, prospects for birth control under capitalism; and second, the probable effectiveness of population restraint (assuming it to be feasible).

For the first issue, there is very little to go on. We do have Engels' famous assertion that "if at some stage communist society finds itself obliged to regulate the production of human beings... it will be precisely this society, and this society alone, which can carry this out without difficulty" (Engels to Kautsky, 1 February 1881; in Ronald Meek, 1971, p. 120). The implied difficulties facing population control under capitalism and the form control would take under communism are unfortunately not clarified. But Marx himself pointed to the steady decline in the rate of population growth over the years—the fact that "although the absolute increase of the English population in the last half century was very great, the relative increase or rate of growth fell constantly" (*Capital*, I, p. 648). And, citing the Registrar General, "rapidly as the population has increased, it has not kept pace with the progress of industry and wealth" (p. 651), evidence of which progress is provided by a wide variety of indexes. Presumably then, he realized that some form or other of population control was being practiced. Notwithstanding, however, the living standards of the masses were, he supposed, falling.[10]

To ascribe the presumed deterioration in real wages to a greater proportional decline in the growth rate of labor demand than supply (as Marx must have done) is open to the objection that, with no check at all to the population growth rate, the deterioration would have been sharper still; while conversely a more determined program of birth control might have been more effective in checking and perhaps halting the deterioration. Here it is pertinent to refer to Marx's condemnation of the "folly of...the economic wisdom that preaches to the labourers the accomodation of their number to the requirements of capital," since "the mechanism of capitalist production and accumulation constantly affects this adjustment" (*Capital*, I, p. 644). This is not a convincing position. The decelerating growth rate of labor demand is an on-going process reflecting exogenous technical progress (the general context of the foregoing comment). A simultaneous program of continuous population control acting on the growth rate of labor supply must tend (at the least) to ease the downward pressure on real wages.

---

[10] Marx alluded favorably to Gladstone's recognition in 1843 of falling real wages—"an increase in the privations and distress of the labouring class and operatives ...[despite] a constant accumulation of wealth in the upper classes, and a constant increase of capital"; but denied Gladstone's assertion in 1863 that "while the rich have been growing richer, the poor have been growing less poor" (*Capital*, I, 651–2). Gladstone's case turned on the lower prices of wage goods which Marx rejected on the basis of data relating to meat, butter, milk, sugar, salt, coals "and a number of other necessary means of subsistence." Similarly, amenity—with special reference to house room—tended to deteriorate, for which proposition extensive evidence is brought (pp. 657f). For a catalogue of references to Marx's allusions to a contemporary fall in living standards in Britain, see Gottheil (pp. 157f). Before his working-class audience of 1865 Marx was more circumspect: "...many contemporary writers have wondered that English capital having grown in the last twenty years so much quicker than English population, *wages should not have been more enhanced*" (*Marx-Engels*, I, p. 445; emphasis added)—to which the "progressive change in the composition of capital" is given in reply.

Marx's case against the Malthusians was reinforced by allusion to Ireland (1840–60), where notwithstanding actual population decline, real wages had failed to improve—indeed, had fallen (p. 706):

> Here, then, under our own eyes and on a large scale, a process is revealed, than which nothing more excellent could be wished for by orthodox economy for the support of its dogma: that misery springs from absolute surplus-population, and that equilibrium is re-established by depopulation....
>
> What were the consequences for the Irish labourers left behind and freed from the surplus-population? That the relative surplus-population is today as great as before 1846; that wages are just as low, that the oppression of the labourers has increased, that misery is forcing the country towards a new crisis. The revolution in agriculture has kept pace with emigration. The production of relative surplus-population has more than kept pace with the absolute depopulation. [pp. 703–04]

But this too is unconvincing, for orthodox classical economics did not constitute a predictive engine allowing a forecast of the course of wages. The intervention of "disturbing causes," in the present case the agricultural revolution involving consolidation of farms, conversion of arable into pasture and use of machinery, does not contradict the general principle that control of population growth would contribute to the maintenance of wages. Population control was considered a necessary but not always a sufficient condition to assure against falling wages.

An objection to population control as an effective means for positively raising wages is stated in *Theories of Surplus Value* in terms of substitution against labor as an implication of the analysis by Barton and Ricardo:

> Essentially... the whole of the absurd theory of population was thus overthrown, in particular also the claptrap of the vulgar economists, that the workers must strive to keep their multiplication below the standard of the accumulation of capital. The opposite follows from Barton's and Ricardo's presentation, namely that to keep down the labouring population, thus diminishing the supply of labour, and, consequently, raising its price, would only *accelerate* the application of machinery, the conversion of circulating into fixed capital, and hence, make the population artificially "redundant".... [Part II, p. 578]

This same argument is made also in *Capital*, in the analysis of *cyclical* fluctuations discussed above, to explain the rejection of the role of absolute variation of the population:

> Yet this is the dogma of the economists. According to them, wages rise in consequence of accumulation of capital. The higher wages stimulate the working population to more rapid multiplication, and this goes on until the labour-market becomes too full, and therefore capital, relatively to the supply of labour, becomes insufficient. Wages fall, and now we have the reverse of the medal. The working population is little by little decimated as a result of the fall in wages, so that capital is again in excess relatively to them.... Then comes again the time, when the supply of labour is less than the demand, wages rise, and so on. A beautiful mode of motion this for developed capitalist production! Before, in consequence of the rise of wages, any positive increase of population really fit for work could occur, the time would have been passed again and again, during which the industrial campaign must have been carried through, the battle fought and won. [I, pp. 637–38]

Marx's objections are then applied to the increase in real wages in English agricultural districts following a sudden and unusual fall in local labor supplies (1849–59):

> What did the farmers do now? Did they wait until, in consequence of this brilliant remuneration, the agricultural labourers had so increased and multi-

> plied that their wages must fall again, as prescribed by the dogmatic economic brain? They introduced more machinery, and in a moment the labourers were redundant again in a proportion satisfactory even to the farmers. There was now "more capital" laid out in agriculture than before, and in a more productive form. With this the demand for labour fell, not only relatively, but absolutely. [p. 638]

Two points must here be made. First, substitution against labor in the event of an absolute fall in labor supply of the foregoing kind could only limit the resultant wage increase; there is no reason to expect the creation of an excess labor supply with downward pressure on the wage.[11] Second, the downward secular pressure for Marx reflects a continuous decline in the growth rate of labor demand (due to exogenous conversion) relative to that of labor supply; and Marx has given us no reason why population control would not, at the least, check that decline.

It has been well said that Marx's purpose was "to establish that capital produced an industrial reserve army no matter what the supply of labour power" so that "we could explain poverty and unemployment without reference to the processes of social reproduction that were frequently invoked though poorly understood by the classical political economists" (Harvey, pp. 163–64). But this purpose he did not achieve because the secular generation of excess labor supplies in his model *required* absolute population increase considering the presumed net increase in labor demand; and because nowhere is it convincingly explained why (*ceteris paribus*) a reduced population growth rate cannot prevent or retard the falling wage trend as Malthus and the orthodox economists maintained. In brief, Marx failed to justify what Sweezy has called the "paradoxical effect of creating unemployment" by "slowing down the rate of population growth" (p. 223).

---

[11] The same objection applied against Barton's analysis; cf. my 1971 article, pp. 112, 124.

REFERENCES

Baumol, William J., "Marx and the Iron Law of Wages," *American Economic Review Proceedings*, May 1983, *73*, 303–08.

Gottheil, Fred M., *Marx's Economic Predictions*, Evanston: Northwestern University Press, 1966.

Harvey, David, *The Limits to Capital*, Oxford: Basil Blackwell, 1982.

Hollander, Samuel, "The Development of Ricardo's Position on Machinery," *History of Political Economy*, Spring 1971, *3*, 105–35.

______, "The Wage Path in Classical Growth Models: Ricardo, Malthus and Mill," *Oxford Economic Papers*, 1984, forthcoming.

______, "'Dynamic Equilibrium' with Constant Wages: J. S. Mill's 'Malthusian Analysis of the Secular Wage Path'," *Kyklos*, 1984, forthcoming.

Levy, David, "Some Normative Aspects of the Malthusian Controversy," *History of Political Economy*, Summer 1978, *10*, 271–85.

Mandel, Ernest, *The Formation of the Economic Thought of Karl Marx*, New York; London: Monthly Review Press, 1971.

Malthus, T. R., *An Essay on the Principle of Population*, reprinted from the last edition revised by the author (1826), London: Ward, Locke & Company, 1890.

Marx, Karl, "Wages, Price and Profit," (1865) in *Marx-Engels: Selected Works*, Moscow: Foreign Languages Publishing House, Vol. I, 1958, 398–447.

______, *Capital*, Moscow: Foreign Languages Publishing House, Vol. I, 1954; Vol. II, 1968; Vol. III, 1962.

______, *Theories of Surplus Value*, Moscow: Progress Publishers, Part II, 1968.

______, *Grundrisse*, Harmondsworth: Penguin Books, 1973.

Meek, Ronald L., *Marx and Engels on the Population Bomb*, 2d ed., New York: Ramparts Press, 1971.

Morishima, Michio, *Marx's Economics: A Dual Theory of Value and Growth*, Cambridge: Cambridge University Press, 1973.

Padover, Saul K., *The Letters of Karl Marx*, Englewood Cliffs: Prentice-Hall, 1979.

Ricardo, David, *Principles of Political Economy* *and Taxation*, in P. Sraffa, ed., Vol. I, *The Works and Correspondence of David Ricardo*, Cambridge: Cambridge University Press, 1951.

Rosdolsky, Roman, *The Making of Marx's 'Capital'*, London: Pluto Press, 1977.

Seccombe, Wally, "Marxism and Demography," *New Left Review*, January-February, 1983, No. 137, 22–47.

Sweezy, Paul M., *The Theory of Capitalist Development*, New York: Monthly Review Press, 1956.

Tucker, G. S. L., "Ricardo and Marx," *Economica*, August 1961, *28*, 252–69.

Regional Productivity Growth in U.S. Manufacturing: 1951–78

By CHARLES R. HULTEN AND ROBERT M. SCHWAB*

Two issues have dominated the discussion of U.S. economic growth over the last decade: the dramatic slowdown in national productivity growth, and the economic decline of the older regions of the United States.[1] Each of these issues has been studied separately, but there have been few attempts to examine them jointly. Yet, there are good reasons to believe the two issues are connected. The older Snow Belt region of the United States has frequently been characterized as an area whose economic performance has been slowed by an aging public infrastructure, deteriorating urban environment, an obsolete capital stock, and what Mancur Olson (1983a) has termed "institutional sclerosis." If this characterization is correct, it could substantially explain both the overall slowdown in national productivity, and the variation in growth rates across regions.

This paper addresses these issues using the Denison-Kendrick-Jorgenson and Griliches sources of growth framework. The growth rate of real value-added in total manufacturing is disaggregated by nine census divisions and then allocated among the growth rates of regional capital stock, regional labor input, and regional total factor productivity. The results for the 1951–78 period are then used to investigate the causes of differential regional growth and the location and source of the slowdown in manufacturing productivity.

The results of our regional sources of growth analysis are summarized in Table 1. Real value-added in the manufacturing industries of the Sun Belt grew at twice the rate of the Snow Belt, and different factors account for the growth of the two regions.[2] Table 1 also suggests that the regional variation in output growth was not due to differences in productivity growth, but instead to variation in the rates of growth of capital and labor. We thus find no evidence to support the hypothesis that an aging public infrastructure, obsolete capital stock or higher rate of unionization have slowed total factor productivity growth in the Snow Belt. The conclusion is reinforced by our finding that labor productivity (output per worker) grew at roughly the same rate in both regions.

The results underlying Table 1 can be used to analyze the slowdown in the growth of labor productivity. Our basic finding is that the slowdown in U.S. manufacturing labor productivity was ubiquitous: growth slowed in eight of the nine census divisions between the 1965–73 and 1973–78 periods. The slowdown was in fact slightly greater in the

---

*The Urban Institute, 2100 M Street, NW, Washington, D.C. 20037, and the Bureau of Labor Statistics; and University of Maryland, College Park, MD 20742, and The Urban Institute, respectively. We thank Janice McCallum, Leland Deck, and Carolyn Estey for their valuable assistance. We also gratefully acknowledge the assistance of Jerome Mark, Michael J. Harper, and Kent Kunze of the Bureau of Labor Statistics in providing unpublished data. We thank Frank de Leeuw, Charles Brown, James Medoff, Mancur Olson, and George Peterson for helpful comments and suggestions. This research was supported by the Office of Community Development, U.S. Department of Housing and Urban Development, the Office of Tax Analysis, U.S. Department of the Treasury, and a Sloan Foundation grant to the University of Maryland.

[1] The decade of the 1970's was a period of unusually slow economic growth in the United States. For private nonresidential business as a whole, J. R. Norsworthy, Michael Harper, and Kent Kunze (1979) have estimated that growth in labor productivity fell from an average annual rate of 3.2 percent over the period 1948–65, to 2.3 percent over the period 1965–73, and to 1.1 percent for 1973–78. Since then, productivity growth has all but ceased. Various explanations of this slowdown have been explored, but as yet no consensus has emerged on the relative importance of each of these factors. See Michael Darby (1982), Edward Denison (1979), John Kendrick (1980), Ernst Berndt (1980), and other papers in Federal Reserve Bank of Boston for a review of the various factors underlying the productivity slowdown.

[2] We define the Snow Belt as the New England, Middle Atlantic, East North Central, and West North Central census divisions. The Sun Belt includes the South Atlantic, East South Central, West South Central, Mountain, and Pacific divisions.

TABLE 1—THE REGIONAL SOURCES OF GROWTH IN U.S. MANUFACTURING, 1951–78

Average Annual Growth Rate of:	Snow Belt	Sun Belt	U.S. Total Manufacturing
Real Value-Added	2.45	4.94	3.31
Capital Stock,			
Weighted by Capital's	0.62	1.63	0.96
Share of Value-Added	(25.3)	(33.1)	(29.0)
Labor Man-Hours,			
Weighted by Labor's	0.03	1.69	0.65
Share of Value-Added	(1.2)	(34.3)	(19.6)
Total Factor	1.80	1.61	1.70
Productivity	(73.5)	(32.7)	(51.4)

Source: See text.
Notes: The percentage contributions to output growth accounted for by each source of growth are shown in parentheses. Detail may not add to total due to rounding.

Sun Belt than in the Snow Belt. What small differences in the slowdown across regions we do observe can be attributed entirely to differences in the growth of the capital-labor ratio, rather than to the growth of total factor productivity.

The remaining sections have the following organization. The theoretical model underlying the sources of growth framework is set out in Section I, and the data used in the calculations are described in Section II. Section III sets out the results of the regional sources of growth analysis, and Section IV presents results on the productivity slowdown.

I. Theoretical Framework

The sources of growth framework used in this paper is derived from the neoclassical theory of production. A Hicks-neutral production function is assumed to exist which relates real product to real factor input, and factors are assumed to be paid the value of their marginal products. The growth rate of real product is partitioned into the share-weighted growth rates of the inputs (interpreted as a movement along the production function), and a residual category associated with total factor productivity (interpreted as the shift in the production function).

In relating this framework to aggregate manufacturing, we assume gross output Q_t is a function of capital K_t, labor L_t, intermediate input M_t, and that the technology is weakly separable between value-added and intermediate goods. We also assume that technical change is value-added augmenting. Thus,

$$(1) \qquad Q_t = F(A_t V(K_t, L_t), M_t).$$

Under profit maximization, the value of marginal product condition implies

$$(2) \quad \partial Q_t / \partial M_t = p_t^M / p_t, \quad \partial Q_t / \partial K_t = c_t / p_t,$$
$$\partial Q_t / \partial L_t = w_t / p_t,$$

where p_t is the price of output, and c_t, w_t, and p_t^M are the relevant factor prices. Logarithmic differentiation of (1), combined with (2), yields the Divisia sources of growth equation:

$$(3) \qquad \dot{v}_t = S_K^t \dot{k}_t + S_L^t \dot{l}_t + \dot{a}_t,$$

where small letters with dots denote growth rates (for example, $\dot{q}_t = (d \log Q_t / dt)$), and S_K^t and S_L^t are the factor's shares in value-added Y_t (i.e., $c_t K_t + w_t L_t$):

$$(4) \qquad S_K^t = c_t K_t / Y_t$$
$$\text{and} \qquad S_L^t = w_t L_t / Y_t.$$

Imposing constant returns to scale on $V(\cdot)$ implies that $S_K^t + S_L^t = 1$.

All magnitudes in (3) are observable except for $\dot{a}_t$, the growth rate of total factor productivity. This variable can thus be inferred as a residual, a procedure which implies that $\dot{a}_t$ combines measurement error with the intrinsic efficiency change that $\dot{a}_t$ is supposed to measure.

In practice, the continuous growth rates and factor shares of (3) must be replaced by their discrete time counterparts. This is generally accomplished by using logarithmic differentials (for example, $\dot{q}_t = (\log Q_t - \log Q_{t-1})$) in place of the continuous growth rates, and by using average shares in place of (4):

$$(5)\qquad \bar{S}_K^t = \left(S_K^t + S_K^{t-1}\right)/2,$$

$$\bar{S}_L^t = \left(S_L^t + S_L^{t-1}\right)/2.$$

The resulting discrete version of the Divisia index is termed, alternatively, the translog index or the Tornqvist index. W. Erwin Diewert (1976) has shown that this discrete index yields exact results if the underlying production function has the translog form.

Our choice of the value-added specification of technology (1) instead of the more general gross output specification, $Q_t = B_t F(K_t, L_t, M_t)$, is dictated by data constraints.[3] Our primary sources of data are the *Census of Manufacturers* and the *Annual Survey of Manufacturers*, which present estimates of intermediate goods in a form that makes the gross output approach difficult to implement. First, purchased services are not treated as intermediate inputs, and are included in a residual category along with income accruing to capital. Second, the manufactured goods produced in one region and used in another region as intermediate inputs are not netted out, so that the sum of the gross products (Q_t) for all regions is not equal to gross product at higher levels of aggregation. As noted in the paper by Hulten (1978), this makes it difficult to compare the nine census divisions with the Sun Belt-Snow Belt aggregates, or with the national aggregates.

In order to minimize these problems, we have chosen to reconcile our census-based estimates with the official estimates of the Bureau of Labor Statistics (BLS) multifactor productivity program. The BLS estimates for manufacturing are based on the value-added specification (1), and we have thus adopted this assumption for our regional analysis.

TABLE 2—SELECTED PRICE AND QUANTITY RATIOS AND OTHER DATA, 1951–78

	Snow Belt	Sun Belt	U.S. Total Manufacturing
Price Ratios:			
Wage-Output Price	4.34	3.82	4.15
Wage-Rent	20.43	16.47	18.83
Quantity Ratios:			
Real Value-Added-Labor	5.72	5.57	5.68
Capital-Labor	7.60	8.59	7.97
Other Data:[a]			
Rate of Return	14.2	15.9	14.9
Regional Share of Value-Added	65.1	34.9	100.0
Labor's Shares of Regional Value-Added	72.4	65.4	69.9

Source: See text.

[a] Shown in percent.

II. Data

This section summarizes the data used in this study. An overview of these data for the broad Sun Belt and Snow Belt regions is presented in Table 2. In addition, since there is substantial variation across divisions, we have included similar data for each of the four Snow Belt divisions and each of the five Sun Belt divisions (for 1951–78 as well as three subperiods, 1951–65, 1965–73, and 1973–78) in Table 3.[4]

A. *Output*

Our basic regional series for value-added in current dollars is obtained from the *Census of Manufacturers* (*CM*) and *Annual Survey of Manufacturers* (*ASM*) for the years 1951–78. However, because value of intermediate services is not subtracted from value of shipments to arrive at valued-added, census value-added is overstated. To address this problem, we have adjusted regional value-added from census to reflect the BLS estimates of national value-added for manufacturing. Correspondingly, the components of regional value-added—labor compensation and capital income—were scaled to the BLS national totals.

---

[3] The value-added and gross output specifications are alternative special cases of a more general technology. Norsworthy and David Malmquist (1982) have undertaken econometric tests of these restrictions using data similar to those used in this study. Those tests rejected the value-added specification in favor of the gross output specification.

[4] A more complete description of the data is available from the authors upon request.

Regional output deflators were not available from *CM-ASM* or from any other source. We therefore used a national deflator obtained from BLS which is based on the U.S. *National Income and Product Accounts* (*NIPA*).

B. *Labor*

The *CM-ASM* contain regional data on the nominal dollar value of wages for production and nonproduction workers separately. These sources also contain the hours worked and employment of production workers, but only the employment of nonproduction workers. The latter was converted to an estimate of hours worked by multiplying by 2,080, the standard number of hours worked during the year assuming a 40-hour work week. The growth of our Divisia index of labor input represents the growth of production and nonproduction workers weighted by each component's share of total labor income.

C. *Capital*

We have estimated the price and quantity of services from plant, equipment, inventories, and land for each region. Our estimates of regional stocks of plant and equipment are based on the perpetual inventory method.[5] This method requires a time-series on gross investment in constant dollars, a benchmark for the capital stock, and an estimate of economic depreciation. The *CM-ASM* provide regional investment data on nominal dollar investment in plant and equipment separately for the 1951–78 period. Our 1929–50 data are based on Patrick Huntley (1961).[6]

We assume that since equipment is bought and sold in national markets, the equipment deflator is the same in all regions. Nominal dollar investment in equipment is therefore converted to constant dollars with BLS/BEA equipment deflators.

Since structures are purchased in regional markets, however, we developed regional structures deflators based on the Boeckh construction cost index for industrial structures for a set of twenty cities and the BLS/BEA national structures deflator. These regional indexes are equivalent to deviations from the BLS national average.

Hulten and Frank Wykoff (1981a, b) show that the assumption of constant economic depreciation is a satisfactory approximation for the estimation of perpetual inventory stocks of capital. The depreciation rate for structures (3.61 percent per year) and for equipment (14.64 percent per year) used in this paper are weighted averages of the depreciation rates for various assets estimated by Hulten and Wykoff, where the weights reflect the mix of assets in the manufacturing sector.

Since our investment series for equipment goes back to 1929, and since the rate of depreciation exceeds 14 percent, only 3 percent of the 1929 benchmark remains in 1951 (the beginning year of this study). Little accuracy is thus lost by assuming the equipment benchmark to be zero. The structures benchmark is another matter, since the depreciation rate is 3.61 percent, and 45 percent of the 1929 benchmark survives to 1951. We therefore impute a 1929 regional benchmark for structures by allocating the BLS all-manufacturing stock of structures to the nine divisions in proportion to 1929 regional manufacturing employment.

Inventory data by region are obtained directly from *CM-ASM* for some years and deflated to constant dollars using the all manufacturing *NIPA* deflator. For the years 1951 to 1961, 1964, and 1965, the stock of inventories was estimated by assuming that the ratio of regional value added to inventory was equal across divisions. The inventory stocks were then scaled to BLS national totals.

The stock of land by division is not available from any source. We estimated regional stocks by allocating the BLS all-manufacturing total to each division according to that

---

[5] See Allan Young and John Musgrave (1980) for a general discussion of the perpetual inventory method and of the problems associated with the measurement of capital stock.

[6] We have used the biproportional matrix method to divide Huntley's estimates of total regional gross investment in manufacturing into separate estimates of regional investment in plant and equipment.

TABLE 3—SELECTED PRICE AND QUANTITY RATIOS AND OTHER DATA

	NE	MA	ENC	WNC	SA	ESC	WSC	M	P	U.S. Total
Rate of Return										
1951–65	.121	.128	.126	.164	.146	.165	.139	.141	.134	.134
1965–73	.149	.157	.153	.207	.177	.215	.177	.173	.152	.165
1973–78	.147	.160	.148	.214	.176	.210	.207	.192	.191	.175
1951–78	.132	.141	.136	.184	.158	.183	.161	.158	.148	.149
Regional Share of Value-Added										
1951–65	.076	.244	.305	.060	.100	.044	.053	.015	.103	1.000
1965–73	.068	.212	.286	.065	.117	.056	.065	.019	.113	1.000
1973–78	.062	.182	.275	.070	.124	.061	.084	.023	.120	1.000
1951–78	.071	.223	.294	.063	.109	.050	.063	.018	.109	1.000
Labor's Share of Regional Value-Added										
1951–65	.749	.728	.698	.664	.639	.608	.569	.638	.712	.691
1965–73	.755	.740	.720	.666	.661	.601	.582	.661	.720	.699
1973–78	.791	.751	.765	.683	.697	.649	.567	.679	.715	.718
1951–78	.760	.736	.719	.668	.658	.615	.573	.653	.716	.699
Wage-Output Price										
1951–65	3.23	3.52	3.71	3.30	2.73	2.72	3.05	3.40	3.81	3.41
1965–73	4.40	4.76	5.05	4.56	3.75	3.69	4.14	4.50	5.11	4.59
1973–78	5.26	5.60	6.21	5.55	4.74	4.60	5.09	5.38	5.97	5.54
1951–78	3.95	4.27	4.57	4.08	3.41	3.36	3.76	4.09	4.59	4.15
Wage-Rent										
1951–65	15.23	16.77	17.28	12.23	11.56	10.38	13.60	14.62	16.62	15.23
1965–73	21.12	21.58	23.94	15.50	15.89	13.03	18.03	18.73	23.43	20.06
1973–78	30.19	28.29	34.74	19.86	22.85	18.69	20.93	24.03	26.14	26.02
1951–78	20.13	20.43	22.78	14.68	15.17	12.89	16.31	17.72	20.61	18.83
Real Value-Added-Labor										
1951–65	4.27	4.80	5.26	4.91	4.23	4.43	5.31	5.28	5.31	4.88
1965–73	5.78	6.39	6.96	6.77	5.62	6.08	7.06	6.76	7.04	6.51
1973–78	6.65	7.43	8.12	8.12	6.79	7.08	8.97	7.93	8.35	7.70
1951–78	4.97	5.56	6.07	5.79	4.93	5.19	6.23	6.04	6.15	5.67
Capital-Labor										
1951–65	5.16	6.21	7.43	6.20	6.52	6.68	10.24	8.16	6.66	6.79
1965–73	6.69	7.48	9.11	7.71	8.04	8.59	12.78	9.55	9.08	8.53
1973–78	7.74	9.18	10.45	9.18	9.76	10.01	15.95	11.35	10.32	10.20
1951–78	6.11	7.16	8.52	7.23	7.62	7.92	12.10	9.20	8.09	7.97

Source: See text. NE: New England; MA: Middle Atlantic; ENC: East North Central; WNC: West North Central; SA: South Atlantic; ESC: East South Central; WSC: West South Central; M: Mountain; P: Pacific.

division's share of the total structure stock. A separate deflator was estimated for each division using a procedure similar to that used for estimating the regional structure deflators. Regional variation in the BLS land deflator was introduced using Federal Housing Administration regional land prices.

The rental price (the cost of using capital services for one year) must also be imputed for each of these four types of capital. For structures and equipment, we use the rental price formulation in Robert Hall and Dale Jorgenson (1967):

$$(6)\quad c=\frac{1-\tau z-\theta+\beta}{1-\tau}\left((r+\delta)q-\dot{q}\right),$$

where τ is the maximum statutory corporate tax rate, z is the present value of depreciation deductions for tax purposes per dollar of investment, θ is the rate of the investment tax credit, β is the Long Amendment parameter ($\beta=\tau\theta z$ in 1962 and 1963; zero in other years), and r is the nominal rate of return on capital. The tax parameters are based on Hulten, James Robertson, and Sally Davies (1981), and the depreciation parameter δ is the same used in the perpetual inventory calculation. The procedure used to estimate r is discussed below.

The rental prices of inventories and land were estimated from a variant of (6) in which the rate of depreciation was assumed to be

Table 4—Sources of Growth in Real Value-Added and Labor Productivity 1951–78 (Percentage growth)

	Growth of Real Value-Added				Growth of Labor Productivity		
	Real Value-Added	Capital[a]	Labor[b]	*TFP*	Labor Productivity	Capital-Labor[a]	*TFP*
New England	2.24	0.48	−0.06	1.82	2.26	0.45	1.82
Middle Atlantic	1.78	0.44	−0.35	1.70	2.25	0.56	1.70
East North Central	2.62	0.68	0.14	1.79	2.34	0.55	1.79
West North Central	4.19	1.20	0.97	2.02	2.70	0.68	2.02
South Atlantic	4.49	1.42	1.39	1.69	2.31	0.62	1.69
East South Central	5.09	1.90	1.58	1.61	2.45	0.83	1.61
West South Central	5.59	2.40	1.77	1.42	2.45	1.03	1.42
Mountain	5.87	1.90	2.58	1.39	1.91	0.52	1.39
Pacific	4.76	1.26	1.84	1.67	2.15	0.48	1.67
Total	3.31	0.96	0.65	1.70	2.33	0.63	1.70
Snow Belt	2.45	0.62	0.03	1.80	2.36	0.56	1.80
Sun Belt	4.94	1.63	1.69	1.61	2.31	0.70	1.61

Note: Detail may not add to total due to rounding.
[a] Weighted by capital's share of value-added.
[b] Weighted by labor's share of value-added.

zero, the marginal tax rate equalled the maximum corporate marginal rate, and the investment credit was zero.

We assume that all assets within a division earn the same after-tax rate of return. We therefore solve for that r such that, given our estimates of each of the four regional stocks of capital, the total rentals exhaust regional capital income; by definition, capital income is equal to value-added less wages.

Our estimates of the growth of total capital input are based on the Divisia indexing procedure. The growth rates of plant, equipment, inventories, and land are weighted and summed to obtain the index of regional aggregate capital input. The weights used in this calculation are each type of capital's share of regional capital income.

III. The Sources of Regional Growth

The data as described allow us to allocate the growth of real value-added among the various sources of growth. This allocation follows equation (3), and is reported in Table 4 in the form of average compound growth rates for the period 1951–78. The corresponding sources of growth allocation for the subperiods 1958–65, 1965–73, and 1973–78 are reported in Section IV where we discuss the productivity slowdown.

It is immediately evident from the left-hand panel of Table 4 that output grew faster in the new regions of the Sun Belt than the older Snow Belt regions. The former grew at twice the rate of the latter, and none of the Snow Belt divisions exceeded any of those in the Sun Belt. The sources of growth were also quite different in the two areas. In the Snow Belt, total factor productivity accounted for nearly three-quarters of the growth in manufacturing real value-added and capital accumulation accounted for virtually the entire remainder. In the Sun Belt, however, each source accounted for one-third of total growth, implying a smaller role for total factor productivity and a much larger role for labor than in the Snow Belt. And, while the growth rate of capital is much larger in the Sun Belt, its relative contribution to growth is only slightly larger.

Nationally, total factor productivity was the dominant source of growth in manufacturing, with capital formation a distant second, and labor growth a clear last. The relative importance of total factor productivity *over time* should not, however, be mistaken for its importance in explaining cross-sectional differences among the regions. It is indeed evident from Table 4 that the growth rate of total factor productivity was slightly higher in the Snow Belt than in the Sun Belt, and that the variation across regions in the growth of real value-added was due to variations in the growth of capital and labor.[7] In fact, 66.8 percent of the differential growth between the Snow Belt and Sun Belt is explained by the differential growth in share-weighted capital input, and −7.7 percent due to differential total factor productivity growth.[8]

Thus, while input growth was not the main driving force of growth over time, it was the main determinant of regional variation in growth rates. This point is particularly noteworthy because much of the discussion of the economic decline of the Snow Belt has focused on the flight of capital and labor to the newer regions of the United States. Our findings confirm the importance of input migration. They do *not*, however, confirm the hypothesis that it is due to a deterioration in overall efficiency in the older regions.

B. *Labor Productivity*

Another way to make this point is to focus on regional trends in labor productivity. The growth in real value-added per man-hour is shown in the right-hand panel of Table 4, along with its decomposition into (*i*) the share-weighted growth rate of the capital-labor ratio, and (*ii*) the growth rate of total factor productivity.[9] This table reveals that labor productivity grew at virtually the same rate in the Snow Belt as in the Sun Belt. Output grew substantially more slowly in the Snow Belt, but the difference in the growth of labor input across regions was roughly equal to the difference in output growth. The overall result is that labor productivity growth was almost the same in both regions, again suggesting that productivity differentials were not the cause of the economic problems of the Snow Belt.

The sources of labor productivity growth are also shown in Table 4. Total factor productivity was obviously the predominant source of growth in both regions, with increases in capital per man-hour accounting for less than 30 percent of the total. Furthermore, the interregional variation in the capital-labor ratio was only slightly larger than the variation in total factor productivity. Thus, the lack of significant regional variation in total factor productivity growth translates into a lack of significant variation in labor productivity growth.

IV. The Regional Contribution to the Productivity Slowdown

The results of the preceding section refer to the entire 1951–78 period. During this period, however, a significant slowdown occurred in the growth of real value-added. Nationally, average annual growth in output was 3.45 percent for 1951–65, 3.97 percent for 1965–73, and 1.85 percent for 1973–78. Thus, while growth accelerated somewhat between the first and second subperiods, it fell sharply after 1973. Between 1965–73 (Period II) and 1973–78 (Period III), the average rate of growth declined by more than one-half.

A. *Regional Growth*

Two general questions are raised by the slowdown between Periods II and III: do the results of the last section hold for each of these two subperiods, and what factors explain the slowdown? To answer the first question briefly, total factor productivity is the most important source of growth over time in both subperiods, and capital and labor continue to explain almost all cross-sectional variation. Also, labor-productivity growth in the Sun Belt and Snow Belt are basically the same within each subperiod, so

---

[7] In contrast, Richard Boyce and Alphonse Holtmann (1979) do find substantial regional differences in total factor productivity growth within the five two-digit manufacturing industries they studied. Their econometric study, however, is based on a substantially different methodology and covers different industries, suggesting that the results of two studies may not be completely incompatible. It should also be noted that Michael Denny, Melvyn Fuss, and J. D. May (1981) also find substantial differences in productivity growth across the regions of Canada.

[8] The faster rate of capital formation in the Sun Belt is consistent with our finding in Table 2 that capital earned a higher rate of return in the Sun Belt than in the Snow Belt.

[9] This decomposition follows directly from (3) by subtracting the growth rate of labor from both sides of the equation.

TABLE 5—SOURCES OF THE SLOWDOWN OF GROWTH IN REAL VALUE-ADDED AND LABOR PRODUCTIVITY 1965–73 AND 1973–78
(Percentage point decrease)

	Change in Growth of Real Value-Added				Change in Growth of Labor Productivity		
	Real Value-Added	Capital[a]	Labor[b]	*TFP*	Labor Productivity	Capital-Labor[a]	*TFP*
New England	0.66	0.37	−0.76	1.05	1.68	0.63	1.05
Middle Atlantic	2.43	0.21	1.03	1.20	0.92	−0.28	1.20
East North Central	2.20	0.24	1.02	0.94	0.88	−0.07	0.94
West North Central	2.67	0.08	0.37	2.24	2.12	−0.12	2.24
South Atlantic	3.42	0.79	1.59	1.03	1.02	−0.01	1.03
East South Central	3.85	0.78	1.70	1.37	1.07	−0.31	1.37
West South Central	−0.08	0.46	0.46	−1.01	−0.82	0.18	−1.01
Mountain	2.83	0.02	0.23	2.58	2.39	−0.18	2.58
Pacific	2.23	0.37	−0.71	2.56	3.27	0.71	2.56
Total	2.12	0.29	0.65	1.16	1.19	0.02	1.16
Snow Belt	2.11	0.23	0.73	1.16	1.13	−0.04	1.16
Sun Belt	2.40	0.55	0.71	1.15	1.34	0.19	1.15

Note: Detail may not add to total due to rounding.
[a]Weighted by capital's share of value added.
[b]Weighted by labor's share of value added.

our general conclusion that productivity differentials are not responsible for differences in regional growth appears to be valid for each subperiod.

The second issue—the sources of the growth slowdown—is addressed in the left-hand panel of Table 5. This table subtracts Period III growth rates from the corresponding Period II rates, so that declining growth rates appear as positive numbers, and allocates the differential growth in real value-added among the differential growth rates of weighted capital and labor and total factor productivity: the underlying growth rates for these periods are shown in Table 6. It is evident from Table 5 that a significant slowdown in growth occurred in all regions but the West South Central, so the slowdown was virtually ubiquitous. The magnitude of the slowdown in the Sun Belt as a whole was 2.40 percentage points, while the Snow Belt slowed by 2.10 percentage points. However, since the growth rate of real value-added was much higher in the Sun Belt during Period II (5.69 vs. 2.97 percent in the Snow Belt), the Sun Belt experienced a much smaller relative decline.

Table 5 also suggests that there was a good deal of variation across divisions within the broader Sun Belt and Snow Belt regions. As noted above, West South Central was the only division where the growth of output accelerated. In sharp contrast, the growth rate of output fell further in the South Atlantic, East South Central, and Mountain divisions than in any of the four Snow Belt divisions.[10]

---

[10] The West South Central results in Table 5 may be misleading for the following reasons: a large segment of the petroleum industry is concentrated in the West South Central division (Arkansas, Louisiana, Oklahoma, and Texas), and it is very difficult to disentangle petroleum refining from other aspects of the petroleum industry which should be properly classified as mining rather than manufacturing activity. It is therefore possible that the rapid rise in oil prices during 1973–78 has led us to overstate the growth in West South Central output, and therefore to overstate the growth in total factor productivity which is measured as a residual, since we have used the same output deflator for all divisions. If we redefine the Sun Belt to exclude West South Central, the average annual growth of output fell 3.00 percent in the (now redefined) Sun Belt as compared to 2.11 percent in the Snow Belt between 1965–73 and 1973–78. Similarly, the slowdown in annual total factor productivity growth for the Sun Belt, exclusive of West South Central, is 1.65 percent in contrast to the Snow Belt's 1.16 percent. We interpret this as further evidence that the decline of the Snow Belt cannot be attributed to differences in economic efficiency relative to the Sun Belt.

Table 6—Sources of Growth in U.S. Manufacturing by Region, Average Annual Growth Rates (Percentage growth)

Growth Rate	NE	MA	ENC	WNC	SA	ESC	WSC	M	P	U.S. Total
Real Value-Added										
1951–65	2.30	2.37	2.83	4.16	4.85	5.52	4.98	5.41	5.17	3.45
1965–73	2.42	2.08	3.23	5.25	5.42	6.10	6.22	7.46	5.19	3.97
1973–78	1.76	−0.35	1.03	2.58	2.00	2.25	6.30	4.63	2.96	1.85
1951–78	2.24	1.78	2.62	4.19	4.49	5.09	5.59	5.87	4.76	3.31
Capital[a]										
1951–65	0.38	0.49	0.69	1.06	1.32	1.73	1.89	1.68	1.23	0.85
1965–73	0.74	0.46	0.77	1.39	1.82	2.37	3.11	2.13	1.43	1.19
1973–78	0.37	0.25	0.53	1.31	1.03	1.59	2.65	2.11	1.06	0.90
1951–78	0.48	0.44	0.68	1.20	1.42	1.90	2.40	1.90	1.26	0.96
Labor[a]										
1951–65	−0.12	0.00	0.12	0.77	1.55	1.73	1.46	2.39	2.10	0.66
1965–73	−0.29	−0.33	0.57	1.33	1.82	2.07	2.29	2.88	1.28	0.89
1973–78	0.47	−1.36	−0.45	0.96	0.23	0.37	1.83	2.65	1.99	0.24
1951–78	−0.06	−0.35	0.14	0.97	1.39	1.58	1.77	2.58	1.84	0.65
Total Factor Productivity										
1951–65	2.05	1.89	2.03	2.33	1.97	2.06	1.63	1.34	1.84	1.93
1965–73	1.97	1.96	1.89	2.54	1.78	1.66	0.81	2.44	2.47	1.89
1973–78	0.92	0.76	0.95	0.30	0.75	0.29	1.82	−0.14	−0.09	0.73
1951–78	1.82	1.70	1.79	2.02	1.69	1.61	1.42	1.39	1.67	1.70
Capital-Labor[a]										
1951–65	0.39	0.47	0.57	0.63	0.41	0.59	0.76	0.33	0.35	0.53
1965–73	0.75	0.54	0.50	0.68	0.84	0.98	1.39	0.66	0.89	0.74
1973–78	0.12	0.82	0.57	0.80	0.85	1.29	1.21	0.84	0.18	0.72
1951–78	0.45	0.56	0.55	0.68	0.62	0.83	1.03	0.52	0.48	0.63
Real Value-Added-Labor										
1951–65	2.44	2.36	2.60	2.97	2.39	2.65	2.39	1.67	2.19	2.46
1965–73	2.72	2.50	2.40	3.22	2.62	2.64	2.21	3.10	3.36	2.64
1973–78	1.04	1.58	1.52	1.10	1.60	1.57	3.03	0.71	0.09	1.45
1951–78	2.26	2.25	2.34	2.70	2.31	2.45	2.45	1.91	2.15	2.33

[a] Weighted by share.

Table 5 shows that the drop in the growth of total factor productivity accounted for over one-half of the slowdown at the national level. This is an interesting result in light of our previous finding that total factor productivity growth was generally the most important source of growth over time.

Another significant aspect of Table 5 is the role of capital formation in the slowdown. Nationally, only 14 percent of the slowdown is attributable to a slower rate of capital formation; this figure is 10 and 22 percent for the Snow Belt and Sun Belt, respectively. The widely held perception that the output decline can be explained by inadequate capital formation thus finds little support in Table 5. It is true, however, that the difference in the percentage declines in value-added growth (2.10 vs. 2.40 percent) is accounted for by the regional differential in capital formation (0.23 vs. 0.55 percent). These differences are, however, too small to be of much significance.

B. *Labor Productivity*

The right-hand panel of Table 5 traces the factors underlying the decline in labor productivity. Again, we find that the decline was experienced everywhere except in the West South Central division. The decline in the Snow Belt was slightly less than in the Sun Belt, but the Snow Belt started from a slightly lower base (2.60 vs. 2.80 percent).

It is also evident that total factor productivity is the overwhelming cause of the labor productivity slowdown, and that almost no role can be assigned to a decline in the weighted capital-labor ratio. However, it is also true that the cross-sectional variation is attributable to differences in the declining growth rate of this ratio although, again, the differences are small.[11]

V. Summary and Conclusion

Three general conclusions emerge from this regional sources of growth analysis. First, the productivity slowdown in U.S. manufacturing was not a regionally isolated phenomenon, but occurred quite broadly across regions. Second, the pattern of growth over time at the national level (including the slowdown) is largely accounted for by variation in the growth of total factor productivity. Third, interregional differences are largely a result of differences in the growth of capital and labor input. We thus find little evidence to support the hypothesis that the economic decline of the Snow Belt was due to differences in economic efficiency relative to the Sun Belt. These results are therefore consistent with Olson (1982; 1983b) who argues that special interest groups in the Northeast and Midwest will inhibit growth by encouraging migration of factors of production, but will not necessarily slow the growth of productivity.

---

[11] Here again, the West South Central may cloud the interpretation of Table 5. In the Sun Belt, redefined to exclude West South Central, the slowdown in the annual average growth of labor productivity was 1.84 percent between 1965–73 and 1973–78.

REFERENCES

Berndt, Ernst R., "Energy Price Increases and the Productivity Slowdown in United States Manufacturing," in *The Decline in Productivity Growth*, Boston: Federal Reserve Bank of Boston, 1980.

Boyce, Richard L. and Holtmann, Alphonse, *A Regional Analysis of Earned Income Patterns in Selected Industries in The United States*, Final Report to the Office of Policy Development and Research, U.S. Department of Housing and Urban Development, Washington 1979.

Darby, Michael R., "The U.S. Productivity Slowdown: A Case of Statistical Myopia," Working Paper No. 1018, National Bureau of Economic Research, 1982.

Denison, Edward F., *Accounting for Slower Economic Growth*, Washington: The Brookings Institution, 1979.

Denny, Michael, Fuss, Melvyn and May, J. D., "Intertemporal Changes in Regional Productivity in Canadian Manufacturing," *Canadian Journal of Economics*, August 1981, *14*, 390–408.

Diewert, W. Erwin, "Exact and Superlative Index Numbers," *Journal of Econometrics*, May 1976, *4*, 115–46.

Hall, Robert E. and Jorgenson, Dale W., "Tax Policy and Investment Behavior," *American Economic Review*, June 1967, *57*, 391–414.

Hulten, Charles R., "Growth Accounting with Intermediate Inputs," *Review of Economic Studies*, October 1978, *45*, 511–18.

______ **and Wykoff, Frank C.**, (1981a) "The Estimation of Economic Depreciation Using Vintage Asset Prices: An Application of the Box-Cox Transformation," *Journal of Econometrics*, April 1981, *15*, 367–96.

______ **and** ______, (1981b) "The Measurement of Economic Depreciation," in Charles R. Hulten, ed., *Depreciation, Inflation, and the Taxation of Income From Capital*, Washington: Urban Institute Press, 1981.

______, **Robertson, James W. and Davies, Sally M.**, "Tax Policy Changes and Effective Tax Rates on Income from Capital: 1952–1980," Working Paper, The Urban Institute, 1981.

Huntley, Patrick R., "State Distribution of Manufacturers' Plant and Equipment in Place, 1954–1956," unpublished doctoral dissertation, University of North Carolina-Chapel Hill, 1961.

Jorgenson, Dale W. and Griliches, Zvi, "The Explanation of Productivity Change," *Review of Economic Studies*, July 1967, *34*, 249–83.

Kendrick, John W., "Productivity Trends and the Recent Slowdown: Historical Perspective, Causal Factors, and Policy Options," in William Fellner, ed., *Contemporary Economic Problems*, Washington: American

Enterprise Institute, 1980.

______ **and Grossman, Elliot S.**, *Productivity in the United States: Trends and Cycles*, Baltimore: Johns Hopkins University Press, 1980.

Norsworthy, J. R., Harper, Michael, Jr. and Kunze, Kent, "The Slowdown in Productivity Growth: Analysis of Some Contributing Factors," *Brookings Papers on Economic Activity*, *2*: 1979, 387–421.

______ **and Malmquist, David H.**, "Input Measurement and Productivity Growth in Japanese and U.S. Manufacturing," preliminary working paper, U.S. Bureau of the Census, 1982.

Olson, Mancur, *The Rise and Decline of Nations: Economic Growth, Stagflation, and Social Rigidities*, New Haven: Yale University Press, 1982.

______, (1983a) "The Political Economy of Comparative Growth Rates," in Dennis C. Mueller, ed., *The Political Economy of Growth*, New Haven: Yale University Press, 1983.

______, (1983b) "The South Will Fall Again: The South as Leader and Laggard in Economic Growth", *Southern Economic Journal*, April 1983, *49*, 917–32.

Young, Allan H. and Musgrave, John C., "Estimation of Capital Stock in the United States," in Dan Usher, ed., *The Measurement of Capital*, New York: National Bureau of Economic Research, 1980.

U.S. Department of Commerce, Bureau of the Census, *Annual Survey of Manufacturers*, Washington, various years.

______, *Census of Manufacturers*, Washington, various years.

______, **Bureau of Economic Analysis**, *The National Income and Product Accounts of the United States, 1929–76, Statistical Tables*, Washington 1981.

Confuser Surplus

By EDWARD R. MOREY*

Paul Samuelson noted that consumer's surplus is a topic "...about which an earlier generation of economists were able to indulge in much argumentation" (1947, p. 195). Indulgence persists. The current consumer's surplus literature shows that there is still much for authors to argue about and to confuse readers.[1] This paper begins and ends with what I consider to be the fundamental question for users of the consumer's surplus concept. Do dollar measures of utility changes exist? The answer is seen to hinge critically on what one means by "measures," and on whether one assumes an ordinal or a cardinal preference ordering.

I. Confusing and Seemingly Contradictory Statements: Why?

The evidence on the question, in terms of quotes from the great and almost great appears mixed.[2] In defense of dollar measures of utility changes, Alfred Marshall states

> ...[The consumer] derives from a purchase a surplus of satisfaction. The excess of the price which he would be willing to pay rather than go without the thing, over that which he actually does pay is the economic [money] measure of this surplus of satisfaction. It may be called consumer's surplus.
> [1920, p. 124]

In agreement, John Hicks states "...the best way of looking at consumer's surplus is to regard it as a means of expressing, in terms of money income, the gain which accrues to the consumer as a result of a price fall" (1943, p. 40). Alternatively, Samuelson has a section in *Foundations* entitled, "Why Consumer's Surplus Is Superfluous," in which he finds the concept worthless (1947, pp. 195–97).

More recently, Eugene Silberberg; Donald Richter; Michael Burns; Arnold Harberger; Robert Willig; G. W. McKenzie and I. F. Pearce; and Charles Cicchetti et al. to name a few, have added to the debate. Silberberg states "The simple truth is that there is no unique dollar or money equivalent of a change in utility" (1978, p. 353). In some sort of agreement, Richter notes that "...it is fruitless to try to justify a consumer welfare measure on the grounds that it represents a money index of utility change..." (1977, p. 502). Burns states "we recognize the impossibility of associating a unique evaluation in money terms with the utility difference between any two situations..." (1973, p. 341). Harberger states "That this [i.e., the notion that consumer's surplus should measure changes in utility directly] would be a fruitless pursuit should be obvious..." (1971, p. 788, fn. 3), and most recently, McKenzie and Pearce state "attempts to measure...consumer's surplus...are no more than disguised attempts to measure utility itself. Each can be correctly measured only by measuring utility" (1982, p. 673).

There are, however, other recent statements in the works of these authors and others which suggest that consumer's surplus measures are dollar measures of utility changes. Burns, for example, makes other statements that suggest that the consumer's surplus concept is a useful dollar measure of utility changes. At the beginning of Willig's often-cited article he refers to the compensating and equivalent variation (both measured in dollars) as "...the correct theoretical measures of the welfare impact of changes in prices and income on an individual" (1976, p. 589). He concludes the paper by stating that "...at the level of the individual consumer, cost-benefit welfare analysis can be performed rigorously and unapologetically by means of consumer's surplus" (p. 596). Harberger explicitly suggests that consumer's surplus is a dollar measure of utility change. He states

---

*Department of Economics, University of Colorado, Boulder, CO 80309. I thank M. Bradley, D. Donaldson, P. Graves, U. Kohli, and D. Lehman for helpful suggestions. Thanks go also to the referee whose comments and insights have contributed significantly to the paper. The usual disclaimer applies.

[1] To limit the scope of confusion, I consider only the measurement of utility changes for a utility-maximizing individual in a single time period.

[2] I find a lot of the statements about consumer's surplus seemingly contradictory. Whether they are when interpreted correctly is another matter. The intent of this section is not to suggest that the quoted authors are confused, rather my intent is to show how easily the nonspecialist, like myself, can become confused.

> ...[Equation] (5) in effect converts the change in utility into monetary terms by dividing it by the marginal utility of income. There is obviously no problem when the latter is not changing, but when it does change as a consequence of the action(s) being analyzed, the conversion of utility into money is implicitly carried out at the midpoint of the beginning and ending marginal utilities of income. The criticism that consumer-surplus concepts have validity only when the marginal utility income is constant must be rejected.
>
> [pp. 788–89]

The work of Cicchetti et al. also definitely suggests that dollar measures of benefits (utility) exist. In a nice mix of the theoretical and the applied, they derive a consumer's surplus measure and use it to measure the dollar "benefits" associated with a new ski area. In apparent agreement, McKenzie and Pearce state that "...the equivalent variation ...is an index of utility" (1982, p. 673). It is clear that the "stated" evidence is mixed regarding whether dollar measures of utility changes exist.

Why has this confusion persisted for so long? These seemingly contradictory statements coexist for four reasons:

1) People are often not explicit about whether they believe consumers have *cardinal* or *ordinal* preference orderings.

2) The term *measures* in the expression *dollar measures* of utility changes has not been clearly defined.

3) The usefulness of consumer's surplus measures depends in part on whether one specifies systems of demand equations directly, or whether one begins by specifying the utility function and then deriving the corresponding system of demand equations. Consumer's surplus measures are more useful to advocates of the first approach than they are to advocates of the latter approach, but both groups tend to forget this.

4) The usefulness of the cardinal properties of consumer's surplus measures depends in part on whether one assumes that all aspects of the alternatives were considered or whether one takes the other view that many important components (characteristics, political implications, etc.) are omitted in the calculation of consumer's surplus measures. The cardinal properties are more important to those who hold the latter view.

The following section discusses preferences and measurement while Section III discusses what properties money indices of utility should have and what properties they do have. Section IV takes up the third reason for the confusion, namely that there are different approaches to demand estimation. Section V considers the calculation of money measures. The fourth reason for confusion is discussed in Section VI. The paper closes in Section VII with an example.

II. Preferences and Measurement

This section defines ordinal preferences and cardinal preferences and discusses the measurement of preference with numbers.[3]

Consider the bundle of goods X_i as one element in the set of all possible bundles. Bundles, like all elements in all sets, possess characteristics. One of the characteristics possessed by all bundles, at least from the individual's point of view, is preference (Pareto's "ophelimity"). What we mean by preference depends on whether we attach the adjective ordinal, weakly cardinal or strongly cardinal.

A. *Ordinal Preferences*

Ordinal preference has magnitude; that is, preference for one bundle is capable of being more than, less than, or equal in magnitude to the preference for some other bundle. If given the choice between two bundles (X_i

---

[3] The theory and logic of measurement, like consumer's surplus, is only simplistic when first encountered. See W. S. Torgerson (1958, pp. 1–40) and Bertrand Russell (1938, pp. 157–87).

and X_j), if X_i is chosen (X_iPX_j), then the magnitude of preference associated with X_i is defined to be greater than the magnitude of preference associated with bundle X_j. If the individual is indifferent between X_i and X_j (X_iIX_j), then the bundles are defined to have equal magnitudes of preference. If an individual is capable of ordering the bundles on the basis of preference magnitudes, and if that ordering is complete, reflexive, and transitive, then that individual has at least an ordinal preference ordering.

One can measure preferences that are solely ordinal on a numerical scale. To measure means to associate with each magnitude of preference a symbol such that the relationships between the symbols reflect accurately the corresponding relationships between the different magnitudes of preference. Ordinal magnitudes of preference only order the bundles so the symbols need only reflect that ordering; that is, if X_iPX_j, then the symbol associated with X_i must take precedence over the symbol associated with X_j. Though not conventional, Ernest Phelps Brown (1934) showed that words could be used to measure magnitudes of ordinal preference; for example, if X_iPX_j, this could be reflected by associating the word "duck" with X_i and the word "economist" with X_j since duck takes precedence over economist (it comes first in the dictionary).

Numbers will also work but one must be careful. One can define a utility function $U = \phi(X)$ that associates a number with each bundle such that if X_iPX_j, then $\phi(X_i) > \phi(X_j)$, and if X_iIX_j, then $\phi(X_i) = \phi(X_j)$. All numerical scales have certain properties, but in this case most of these properties are meaningless. The danger is that this is easy to forget. Numbers have names, numbers are ordered, the difference between two numbers is a number (so differences between numbers are ordered), the ratio of two numbers is a number, and the series of numbers has a unique origin indicated by the number "zero." Every numerical scale will have all of these properties regardless of whether they are appropriate for the underlying characteristics being measured. To measure ordinal preferences, one only needs the ordering property of numbers. For this reason, a numerical measurement of ordinal preferences ($\phi(X)$) is unique only up to a monotonic transformation. The function $F(\phi)$, where $\partial F(\phi)/\partial\phi > 0$, defines the class of monotonic transformations of $\phi(X)$.

B. *Weakly Cardinal Preferences*

Assume that, besides being able to rank bundles, the individual is able to ordinally rank the preference differences between bundles. That is, the individual knows whether a move from X_i to X_j is preferred, not preferred, or equally preferred to a move from X_k to X_l. Call this the ability to rank differences. If the individual has this ability, then the difference between two magnitudes of preference is also a magnitude of preference, implying preference has intensity. This is not the case if preferences are solely ordinal. If the ordering of differences is complete, reflexive, and transitive, then preferences are said to be weakly cardinal (Nicholas Georgescu-Roegen, 1968, p. 260).

If one is willing to stipulate some additional axioms, then weakly cardinal preferences can be measured numerically.[4] That is, a function $\phi(X)$ exists such that: 1) $\phi(X_i) \gtreqless \phi(X_j)$ if, respectively, X_iPX_j, X_iIX_j, or X_jPX_i; and 2) $[\phi(X_j)-\phi(X_i)] \gtreqless [\phi(X_l)-\phi(X_k)]$ if, respectively, $(X_i$ to $X_j)P(X_k$ to $X_l)$, $(X_i$ to $X_j)I(X_k$ to $X_l)$, or $(X_k$ to $X_l)P(X_i$ to $X_j)$.[5] To numerically measure weakly cardinal preferences, we need only the ordering property of numbers and the fact that the difference between two numbers is also a number thus is also ordered. For this reason, a numerical measure of weakly cardinal preference ($\phi(X)$) is unique only up to an affine transformation ($\alpha + \beta\phi(X)$ defines the class of affine transformations of $\phi(X)$). At this point one must be careful; the ratio $[\phi(X_j)-\phi(X_i)]/[\phi(X_l)-\phi(X_k)]=r$ is well-defined numerically (it might, for example, be two) but that does not mean that the move from (X_i to X_j) is r times preferred to the move from (X_k to X_l). The ratio of two magnitudes of preference is not defined if preferences are only weakly cardinal. One cannot, therefore, attribute meaning to the ratio properties of the numerical scale.[6]

---

[4] Samuelson (1938) showed that a function $\phi(X)$ that ranks bundles and differences correctly will not exist (Oscar Lange, 1934a, b, thought otherwise) unless some restrictions are imposed on the individual's ordering of differences. One necessary restriction is that if $(X_i$ to $X_j)I(X_l$ to $X_m)$ and if $(X_j$ to $X_k)I(X_m$ to $X_n)$, then $(X_i$ to $X_k)I(X_l$ to $X_n)$. The mathematical function $\phi(X)$ will order differences as if this restriction on preferences is appropriate. If it is not appropriate, the function's ranking of differences will be incorrect.

[5] As Phelps Brown notes, one can also measure weakly cardinal preferences with words (for example, English words to order bundles and French words to order differences).

C. *Strongly Cardinal Preferences*

One must assume that preferences are strongly cardinal before one can meaningfully say that the move from (X_i to X_j) is preferred twice as much as the move from (X_l to X_k). Preferences are strongly cardinal if, in addition to being weakly cardinal, the ratio of two magnitudes of preference differences is also a meaningful magnitude of preference.[7] As with weakly cardinal preferences, the utility function ($\phi(X)$) is unique only up to an affine transformation if preferences are strongly cardinal, but now the ratio $[\phi(X_j)-\phi(X_i)]/[\phi(X_l)-\phi(X_k)]=r$ does reflect a meaningful property of the individual's preferences.

III. What Properties Should a Money Index of Utility Have? What Properties Do they Have?

The answer to the first question depends on whether one believes the consumer has a cardinal preference ordering (weak or strong) or just an ordinal ranking of bundles. The cardinalist would like the money index of utility to be an affine transformation of the utility function. Hence, if the move from (X_i to X_j) provides A utils, the move from (X_l to X_m) provides B utils, and the dollar measure of the change from (X_i to X_j) is \$10, the cardinalist would like the dollar measure of the change from (X_l to X_m) to be (B/A)\$10. That is, the cardinalist would like the dollar index of utility changes to have the same properties as utility.

Alternatively, the ordinalist only requires that the money index rank alternative bundles in the same order the consumer would. The ordinalist only requires that the money index be a strictly monotonic transformation of the utility function. The cardinal magnitude of the difference in utility levels associated with different alternatives can tell one nothing about an ordinalist's intensity of preferences because the ordinalist's preferences do not have intensities.

Concerning what properties money indices do have, is there a money index of utility that is a strictly monotonic transformation of the utility index? The answer is yes. The equivalent variation (EV)[8] will rank bundles in the same order the consumer would. In this sense, dollar measures of utility changes exist. McKenzie and Pearce express this view when they state that "the equivalent variation...is an index of utility" (1982, p. 673). This is what Willig means when he tells us that consumer's surplus is a useful concept.

Is there, however, a money index of utility that is an affine transformation of a cardinally meaningful utility index? That is, does the function $M=\alpha+\beta U$, where α and β are constant scalars, M is money, and U is the cardinal utility index, exist? The answer, in general, is no, as is easy to see.[9] Such a function will exist only if β is a constant scalar that transforms utils into money; β is therefore the inverse of the marginal utility of money and will be a constant only if the marginal utility of money ($\partial U/\partial M$) is constant. The marginal utility of money is not usually assumed to be constant.[10] In this sense, dollar measures of utility changes do not exist. This is what I believe Silberberg, Richter, and Burns meant in the above quotes.

---

[6] Historically, Lange (1934a,b) thought that if the individual can rank bundles and differences then the individual can automatically associate meaning to the ratios of preference magnitudes and he or she can say one move is r times more preferred than some other move. Phelps Brown showed this assertion to be incorrect. This finding is reiterated by Georgescu-Roegen.

[7] A stronger condition comes to mind. One could additionally assume that the ratio of preference magnitudes associated with two bundles (as opposed to the ratio of the preference differences between bundles) is also a meaningful magnitude of preference. One could then say that bundle X_i provides n times as much preference as bundle X_j. Call preferences with this property purely cardinal. In this case, the utility function is only unique up to a linear transformation ($\beta\phi(X)$ defines the class of linear transformations of $\phi(X)$). Purely cardinal preferences are not common in the literature.

[8] The equivalent variation associated with a change from the parametric price-income configuration (P^0, M^0) to the parametric price-income configuration (P', M') is $EV=E[U', P^0]-E[U', P']+M'-M^0$ where $E[\cdot]$ is the expenditure function and U' is the maximum utility level given P' and M'. The value of the EV is invariant to monotonic transformations of the expenditure function. If the indirect utility function has been identified up to a monotonic transformation, then the expenditure function can easily be obtained by solving the indirect function for the level of expenditures.

It is now clear that a money index of utility (i.e., the *EV*) exists which is a strictly monotonic transformation of utility, but remember that the cardinal magnitudes of the *EV*s cannot, in general, tell us anything about an individual's intensity of preference. The fact that the *EV*s are invariant to monotonic transformations of the utility function proves this. Changing a cardinal utility function by a monotonic transformation will change the intensity of preferences, but not the *EV*s.

IV. The Usefulness of Money Measures Depends in Part on How You Estimate Demand Equations

There are two distinct approaches. The first begins by specifying an explicit algebraic form for either the direct or indirect utility function and then derives the corresponding system of demand equations; the second specifies the algebraic form of the demand equations directly. If one adopts the first approach, estimation of the demand equations implies direct knowledge of an estimated utility function. One can then easily use this estimated utility function to ordinally rank projects. Advocates of this approach do not need money measures of utility changes to rank projects and sometimes forget that advocates of the other approach might. This forgetfulness is one of the reasons Samuelson used the adjective "superfluous" (1947, p. 195).

However, if one adopts the second approach, knowledge of the demand functions does not immediately imply knowledge of the ordinal preference ordering. One must solve the integrability problem to obtain the preference ordering. This can be difficult. Advocates of this direct approach to demand estimation are therefore attracted to money measures for their ability to rank projects without direct knowledge of the preference ordering. This group sometimes forgets that the other group does not need these money measures to rank projects.

This forgetfulness probably motivated the following responses to Samuelson's remarks about superfluousness. I. M. D. Little states

> This is a rather surprising passage [referring to Samuelson's remarks]. It implies not only that all individuals are "economic", but also that we have a copy of everyone's preference field filed away.... We can readily agree that consumer's surplus is surplus in formal logic, but the point of the theory surely was to establish a practical criteria.
> [1950, p. 176]

More recently, in the same view, J. M. Currie et al. state

> Samuelson chose to attack the concept [of consumer's surplus] at a purely theoretical level on the assumption that everyone's preference field is known, while the main appeal of the concept ...lies in its relationship with the demand curve. Given that the economist does not have a copy of everyone's preference map filed away, it does not really diminish the practical significance of the concept to argue....
> [1971, pp. 786–87]

---

[9] For a less intuitive proof, see Silberberg's (1972) line integral proof.

[10] Samuelson (1942) has proven that the marginal utility of money cannot be a constant independent of both the income and price levels. However, it can be independent of either the income level or some of the prices. The special case where the marginal utility of money ($\partial U(y,p)/\partial y$) is independent of the income level y is often discussed. The marginal utility of money is independent of y iff $U = V(y,p) = y/e(p)$. Note that $V(y,p) = y/e(p) \Leftrightarrow E(U,p) = Ue(p) \Leftrightarrow U(X)$ is homogeneous of degree one in X. If the price level is constant, M and U will be affine transformations of one another iff $V(y,p) = y/e(p)$. This means that the *EV*s for a group of alternatives will measure the intensity of preferences for those alternatives if the alternatives differ only in terms of the income level and if the individual's cardinal utility function is homogeneous of degree one in X.

As Abram Bergson (1975, p. 38) points out, knowledge of the demand equations (if they fulfill the regularity conditions) is sufficient information, at least in theory, to determine the ordinal preference ordering. One wonders whether Little and Currie et al. do not realize this or whether they just feel that, given the demand equations, it is easier to rank alternatives using consumer's surplus measures than it is to solve the integrability problem directly.

With both approaches to demand estimation, one can only be sure that the ranking of alternatives is correct if one is sure they have deduced the correct demand functions from the market data. On this point, one can never be sure, not even in the world where demand has no stochastic component.

The one approach specifies an algebraic form for the utility function, derives the corresponding system of demand equations, and only then uses data to estimate the coefficients in those demand equations. Imposing an algebraic form on the utility function with complete ignorance of the data amounts unfortunately to imposing a priori restrictions on the preference ordering. These a priori restrictions are embedded in the demand equations. Since their imposition was not based on the data, the estimated demand equations are not necessarily correct and the resulting ranking could be wrong. This is why advocates of this approach often attempt to choose an algebraic form that imposes as few a priori restrictions on the preference ordering as possible while still keeping the form simple enough to econometrically estimate the parameters in the corresponding demand equations. This explains the current popularity of "flexible functional forms" over historically popular forms such as the Bergson and the Linear Expenditure System. Most flexible functional forms (for example, the Translog and Generalized Leontief) are second-order approximations to the true function at some point of approximation. The derived demand equations are therefore first-order approximations. Unfortunately, not even flexible forms are perfectly flexible over the whole range of the data, so restrictions are inevitably imposed on the preference ordering. There is no free lunch when it comes to algebraic forms.

The other approach to demand estimation specifies the algebraic form for the demand equations directly and then uses the market data to estimate their coefficients. While not as obvious, this approach also imposes, perhaps incorrectly, a priori restrictions on the preference ordering. The algebraic form of the demand equations determines the form of the utility function up to a monotonic transformation. Advocates of this approach should attempt to choose, subject to the regularity and simplicity constraints, the most flexible algebraic forms for the demand equations. Again the best we can hope for is algebraic forms that will approximate the true system of demand equations.

Both approaches suffer from the deficiency that they impose a priori restrictions on the preference orderings. This deficiency cannot be avoided, or said differently, we cannot estimate a perfectly flexible system of demand equations with market data. Even in a world where demand has no stochastic component, an estimated demand system (i.e., one that fits all the observations exactly) will not be unique if the number of market observations is finite, as there are an infinite number of functions that pass through a finite number of points. There is no way of knowing that you have the correct one. One cannot even know that one has the correct function in the neighborhood of some chosen point.

V. Using Money Measures to Rank Alternatives without Direct Knowledge of the Preference Ordering

Suppose you have estimated an individual's system of demand equations and are willing to assume it is correct. How can this information be used to determine how the individual would rank a number of price-income configurations? Attempting to solve the integrability problem is not the easiest way to proceed.

We know that *EV*s will rank the alternatives correctly.[11] As noted earlier, *EV*s are easy to determine if the expenditure function or indirect utility function is known up to a monotonic transformation, but this information is now unavailable.

---

[11] Both the compensating variation (CV) and equivalent variation (EV) will rank bundles correctly. However, the CVs associated with the moves from X_i to X_j and from X_i to X_k will not necessarily rank the two moves correctly, whereas the EVs will.

McKenzie and Pearce (1976, 1982) have developed a technique to determine the *EV*s for different alternatives using only the demand functions. Stated simply, they use a Taylor-series expansion to determine the *EV* for a given alternative as a function of the initial price and income level (y^0, p^0), the changes in prices (Δp), and the change in money income (Δy), where all the coefficients in the Taylor-series expansion are first- and higher-order partial derivatives of the demand equations evaluated at (y^0, p^0).

The other major technique for determining an *EV*, when the expenditure function has not been identified, is to approximate it with the Marshallian consumer's surplus measure. Willig argues that "in most applications the error of the approximation will be very small" (p. 589). The McKenzie-Pearce technique does not involve approximation errors if the Taylor-series expansion is evaluated in full. Also note that the McKenzie-Pearce technique does not require complete knowledge of the demand equations, it only requires knowledge of the demand equations in the neighborhood of (y^0, p^0). This gives it a definite informational advantage over the Marshallian measure, which requires knowledge of the demand equations over the complete range of the price variations. The McKenzie-Pearce technique is new, but it should prove quite useful to applied welfare economists.

It has been noted that the cardinal properties of an individual's cardinal preference ordering cannot be inferred from the cardinal properties of a money measure of that ordering. The reader, however, should be aware that this statement can be strengthened. One can never determine the cardinal properties of the individual's cardinal preference ordering (i.e., the individual's intensity of preferences) solely on the basis of the individual's behavior in the market.[12] It cannot be done because any monotonic transformation of the true cardinal utility function will explain behavior just as well as the true function.[13] The true function can only be identified if the individual provides, under interrogation, what Samuelson refers to as "introspective information" (1974, p. 1258) on the intensity of their preferences.

However, since individuals possess introspective information about themselves, they can, after being informed of their own *EV*s for a number of alternatives, use this information to determine their own intensity of preferences for those different alternatives.[14] One uses these dollar amounts in the same way temperature readings can be used to determine how much Palm Springs is preferred at 90°F to Palm Springs at 112°F.

VI. But Don't the Cardinal Properties of the *EV*s Mean Something?

This is a difficult question so before addressing it, let me summarize some pertinent points. 1) Alternatives can be ranked correctly on the basis of their respective *EV*s (a money measure) if all aspects of the alternatives are considered in the determination of those *EV*s. 2) Money measures of changes in utility, such as the *EV*, have cardinal properties independent of whether the underlying preference ordering is cardinal or ordinal. 3) Knowledge of the cardinal differences between *EV*s is not necessary to rank projects. 4) The cardinal magnitudes of *EV*s are invariant to monotonic transformations of the utility function. 5) Knowledge of another person's *EV*s for a number of alternatives tells us nothing about that individual's intensity of preferences for those alternatives. 6) The *EV*s express how much better off each alternative will make us in dollars evaluated at current prices, and these dollar amounts can be compared cardinally in the sense that we can say $100 is twice as much money as $50. 7) If one has cardinal preferences, one can use one's *EV*s for a number of alternatives to determine one's own intensity of preferences for those alternatives. 8) However, you do not need to be informed of your own *EV*s if you know your own preferences and have complete information on all the alternatives.

---

[12] A point known to Pareto; elucidated by Lange (1934a, pp. 224–25), and Roy G. D. Allen (1935, p. 155). This was brought to my attention by the referee.

[13] This makes one wonder. If cardinal preferences don't seem to help us to choose, why have them (if in fact we do have them)? A major thrust of utility theory since Antonelli in 1886, Fisher in 1892, and Pareto, has been to point out that one doesn't need cardinality to explain market behavior. There has been resistance. For a while some welfare economists thought cardinal utility was required to make interpersonal welfare judgements. Bergstrom (1938) showed that they were wrong. More recently, von Neumann and Morgenstern (1964) and others have argued that certain behavior (i.e., gambling) might imply that individuals have cardinal utility. However, the general trend has been to purge the concept from all but principles books. Why then do I imagine, or hope, that my preference ordering is at least weakly cardinal over some range? Maybe one is just better off with cardinal preferences.

[14] One might wonder why an individual would find it useful to have someone else tell them what their own *EV*s are. If one knows their own preferences and has complete information concerning all the alternatives, there is no reason. However, if an individual does not have complete information, he will find *EV*s calculated using complete information useful summary measures.

With these points in mind, the question boils down to how important it is to know that an *EV* of $100 is twice as much as an *EV* of $50. Not everyone agrees on the answer; some take a theoretical perspective while others adopt a more practical point of view.

Stated simply, in the theoretical view the cardinal properties of the *EV* measures have meaning (in the sense of points 6 and 7), but the cardinal information provided is of little use to policymakers.[15] In a world of ordinal preference ordering, all the decision maker needs is a ranking. In a world of cardinal preference ordering, the decision maker can hope for no more than a ranking. The cardinal differences between the *EV*s does not increase the decision maker's ability to rank alternatives and will probably be misinterpreted by the decision maker. If the decision maker desperately wants to know how much one alternative is preferred to another, he or she will be greatly tempted to assume that this desired information is provided by the relative magnitudes of the two alternative *EV*s. Representing this viewpoint, Samuelson brought our integrity into question when, after discussing whether consumer's surplus measures are useful for answering certain welfare questions, he stated "...all of these questions can be more conveniently (and more honestly!) answered in terms of the consumer's ordinal preference field..." (1947, p. 187).

Advocates of the practical view would probably accept the theoretical view if they felt we lived in a world where all aspects of the alternatives were considered when alternatives were ranked, and decisions were made solely on the basis of economic criteria. However, according to the practical view, money measures are quite useful in the real world. Bergson, for example, after summarizing the theoretical view, states "...I don't think I am saying anything that he [Samuelson] would consider very controversial when I hold that in practical work an ordinal ranking may sometimes do, but very often it is helpful, if not indispensible, to know not only more or less but how *much* more or less" (1975, pp. 40–41). Bergson goes on to argue that, in practice, when economists rank alternatives, they often do it on the basis of only a subset of the components of those alternatives: "...the evaluation [of welfare losses and gains] may be incomplete in diverse ways. It may well be incomplete even in respect to some costs and benefits of an economic sort. It would not be surprising, for example, if externalities such as pollution, congestion and the like were often omitted either wholly or in part" (1975, p. 41).

Bergson seems to suggest that although alternative states for an individual are identified in terms of prices, income level, taxes, fixed quantities of certain goods, characteristics, and risk, when economists rank projects, the ranking is done including only some of the components of each of the states (for example, prices).[16] The politician who must choose one of the alternatives knows that the ranking might be wrong because of these omitted factors; therefore, in the words of Bergson,

---

[15] Even if one thinks that knowledge of the cardinal magnitudes of the *EV*s are of little use to policymakers, one would still probably find them personally interesting. Knowing another's *EV*s allows one to determine how much that person's *EV*s, if given to you, would increase your utility. Finding out someone else's *EV* for a project is like hearing that they just won a lottery; there is an automatic tendency to convert the dollar amount into utils on your scale.

[16] My 1982 paper has developed and estimated *EV*s that are a function of prices, characteristics of the goods (or activities), and characteristics of the individual.

> Before commiting himself, the public official will surely wish to know whether, according to any usual calculation, the measure in question would be economically advantageous or disadvantageous; but he might well want to ascertain also whether the gain or loss would be sufficient to make the action desirable or undesirable when due account is taken of costs and benefits beyond the reach of the calculations. True, once such costs and benefits are taken into account, an ordinal ranking is still all that is needed; but in arriving at such a ranking, the public official might be greatly aided if the calculated economic gain or loss were expressed in terms of some cardinally scaled metric. [1975, p. 41]

This argument suggests that the decision maker finds cardinally scaled measures of gains and losses meaningful because we (economists) do not do a very good job of including all the factors. The implication is that our economic ranking would be correct and there would be little reason for the cardinal measure if only we were clever enough to consider all the components of each alternative. Of course, in practice, decision makers do not rank solely on the basis of economic criteria. As Jagdish Bhagwati notes: "Policies are maintained or changed largely for noneconomic reasons; and the (economic) 'cost' involved is a magnitude that is commonly demanded and bandied about in discussions of public policy. Whether we like it or not, this is what the policy makers do want" (1965, p. 213). As Bergson implies, the politician needs cardinally scaled measures of the economic benefits of each alternative to compare them with the political costs of each alternative so as to determine his ordinal ranking of the projects.[17]

VII. A Concluding Example

Closing with an example that demonstrates many of my points, consider three government projects (A, B, C) that cause relative prices to change, and two individuals, Orville the ordinalist and Carleton the cardinalist. Assume we know only their demand equations and, by chance, both systems of demand equations are identical. Using the demand equations and applying McKenzie and Pearce's technique, we determine that the three EVs are $EV_A = \$100$, $EV_B = \$50$, and $EV_C = \$25$.[18]

What can we conclude? For both individuals, we know that project A is preferred to B which is preferred to C, but we don't know anything about how much Carleton prefers A over B or B over C. Orville doesn't think in terms of how much.

The EVs are meaningful in the cardinal sense that \$100 is twice as much money as \$50. Project A makes both individuals \$100 better off at current prices. Project B makes both individuals \$50 better off at current prices. Knowing these amounts and the level of current prices Carleton can figure out how much he prefers project A over B, but there is no way we are going to know unless he tells us. We can't say Carleton likes A twice as much as B.

It is obvious that the politician who must choose one of the three projects would rather know the EVs than just the rankings. Advocates of the practical view would say that the politician needs them so he can compare them to his implicit estimates of the political, and other omitted aspects of the projects. For example, maybe the politician knows that project A will cause an increase in precipitation. Neither Orville nor Carleton likes rain, but the economist did not consider this in ranking (his or her mistake). The politician then weighs the cardinal dollar advantage of A ($50) with the increased rain to try to determine whether Orville's and Carleton's rankings would change if they knew about the rain. The politician might also like to know the EVs because his ranking of the projects will depend on each of their political implications. The $50 advantage of A over B might not be sufficient to overcome the fact that A will decrease his chances of reelection.

---

[17] Cardinally scaled measures could also prove useful if the decision maker must make interpersonal utility comparisons (for example, see if a project has the potential for a Pareto improvement). Interpersonal utility comparisons are important for policy, but for the sake of clarity, this is a topic I have chosen not to discuss.

[18] Before continuing I should note that the McKenzie-Pearce technique is probably the easiest way to get the rankings if one does not get the demand equations by first specifying a preference ordering. However, if we got the demand equation by first assuming a preference ordering, we could have more easily used it to rank the projects and just reported to the policymaker that both individuals prefer A to B and B to C.

Advocates of the theoretical view will be personally interested in knowing Orville's and Carleton's EVs, but will worry that the politician might misinterpret them and conclude that both Orville and Carleton like A twice as much as B.

REFERENCES

Allen, Roy G. D., "A Note on the Determinateness of the Utility Function," *Review of Economic Studies*, February 1935, *2*, 155–58.

______, *Mathematical Analysis for Economists*, London: Macmillan, 1950.

Bergson, Abram (Burk, A.), "A Reformulation of Certain Aspects of Welfare Economics," *Quarterly Journal of Economics*, February 1938, *52*, 310–34.

______, "A Note on Consumer's Surplus," *Journal of Economic Literature*, March 1975, *13*, 38–44.

Bhagwati, Jagdish, "The Pure Theory of International Trade: A Survey," in *Surveys of Economic Theory*, Vol. II, New York: St. Martin's Press, 1965, 156–239.

Burns, Michael E., "A Note on the Concept and Measurement of Consumer's Surplus," *American Economic Review*, June 1973, *63*, 335–44.

Cicchetti, Charles J., Fisher, Anthony C. and Smith, V. Kerry, "An Econometric Evaluation of a Generalized Consumer Surplus Measure: The Mineral King Controversy," *Econometrica*, November 1976, *44*, 1259–76.

Currie, J. M., Murphy, J. A. and Schmitz, A., "The Concept of Economic Surplus and its Use in Economic Analysis," *Economic Journal*, December 1971, *8*, 741–99.

Dupruit, Jules, "On the Measurement of the Utility of Public Works," *International Economic Papers*, 1952, reprinted in *Readings in Welfare Economics*, Homewood: Irwin, 1969; "De l'utilitie et de sa Mesure," 1844, reprinted with comments by M. de Bernardi and L. Einaudi, Turin and Paris, 1934.

Georgescu-Roegen, Nicholas, "Utility," in *International Encyclopedia of the Social Sciences*, New York: MacMillan, 1968.

Harberger, Arnold C., "Three Basic Postulates of Applied Welfare Economics: An Interpretive Essay," *Journal of Economic Literature*, September 1971, *9*, 785–97.

Hause, John C., "The Theory of Welfare Cost Measurement," *Journal of Political Economy*, December 1975, *83*, 1145–82.

Hicks, John R., "The Four Consumer's Surpluses," *Review of Economic Studies*, No. 1, 1943, *11*, 31–41.

______, *Value and Capital*, 2d ed., Oxford: Clarendon Press, 1947.

Lange, Oscar, (1934a) "The Determinateness of the Utility Function," *Review of Economic Studies*, June 1934, *1*, 218–25.

______, (1934b) "Notes on the Determinateness of the Utility Function," *Review of Economic Studies*, October 1934, *2*, 75–77.

Little, I. M. D., *A Critique of Welfare Economics*, Oxford: Clarendon Press, 1950.

Marshall, Alfred, *Principles of Economics*, 8th ed., London: Macmillan, 1920.

McKenzie, G. W., and Pearce, I. F., "Exact Measures of Welfare and the Cost of Living," *Review of Economic Studies*, October 1976, *43*, 465–68.

______ **and** ______, "Welfare Measurement —A Synthesis," *American Economic Review*, September 1982, *72*, 669–82.

Morey, E. R., "Characteristics, Consumer Surplus, and New Activities: Theory and Application," Discussion Paper, Department of Economics, University of Colorado, 1982.

Pearce, I. F., *A Contribution to Demand Analysis*, Oxford: Clarendon Press, 1964.

Phelps Brown, Ernest H., "Notes on the Determinateness of the Utility Function," *Review of Economic Studies*, October 1934, *2*, 66–69.

Richter, Donald K., "Games Pythagoreans Play," *Public Finance Quarterly*, October 1977, *5*, 495–515.

Russell, B., *Principles of Mathematics*, 2d ed., New York: Norton, 1938.

Russell, R. R., and Wilkinson, M., *Microeconomics: A Synthesis of Modern and Neoclassical Theory*, New York: Wiley & Sons, 1979.

Samuelson, Paul A., "The Numerical Representation of Ordered Classifications and the Concept of Utility," *Review of Economic Studies*, October 1938, *6*, 65–70.

______, "Constancy of the Marginal Utility of Income," in Oscar Lange et al., eds., *Studies in Mathematical Economics and Econometrics: In Memory of Henry Schultz*, Chicago: University of Chicago Press, 1942, 75–91.

______, *Foundations of Economic Analysis*, Cambridge: Harvard University Press, 1947.

______, "Complementarity—An Essay on the 40th Anniversary of the Hicks-Allen Revolution in Demand Theory," *Journal of Economic Literature*, December 1974, *12*, 1255–89.

Silberberg, Eugene, "Duality and the Many Consumer's Surpluses," *American Economic Review*, December 1972, *62*, 942–52.

______, *The Structure of Economics: A Mathematical Analysis*, New York: McGraw-Hill, 1978.

von Neumann, J., and Morgenstern, O., *Theory of Games and Economic Behavior*, 3d ed., New York: Wiley & Sons, 1964.

Torgerson, W. S., *Theory and Methods of Scaling*, New York: Wiley, 1958.

Willig, Robert D., "Consumer's Surplus Without Apology," *American Economic Review*, September 1976, *66*, 589–97.

Corn at Interest: The Extent and Cost of Grain Storage in Medieval England

By DONALD N. MCCLOSKEY AND JOHN NASH*

The history of storage of grain in the Middle Ages is important for understanding the past and for its contribution to other studies. It is a simple case of a complex problem in dynamic economics. At the interest rates he faced a medieval farmer seldom stored more than two years in a row; at the transport costs he faced he seldom brought grain from radically different climes. A two-period model and a closed economy are simple conditions for the study of the economics of storage, much simpler than conditions seen nowadays. Further, an assumption about the prevalence of storage underlies much of medieval economic history, and is not irrelevant to modern times. Storage was a species of insurance that could substitute for other species, such as (to pick an example quite at random) scattering of one's holdings of land (McCloskey, 1976). Scattered holdings and the desperate fear of famine they signify are common features of the modern as of the medieval countryside. The force of the economist's argument (see, for example, Theodore Schultz, 1964) that such customs are insurance rather than rural idiocy depends on a quantitative measure of at least one alternative. Storage is the easiest to measure.

The history of storage is important, too, for what can be discovered along the way. Most notably, a measure of the cost of storage sheds light on the prevailing rate of interest, illuminating its hitherto obscure history. The reasoning involved, examined in detail below, is that a store of grain is an investment. Wheat put into storage in October and brought out in November must pay over the month the cost of the barn and the guards, the depreciation of the grain, and the opportunity cost of the funds invested. The opportunity cost is the rate of interest. The rate was shockingly high. Stores of grain were therefore very low, and medieval men lived from hand to mouth—as one might have judged as much from their poetry as from their markets. In the sixteenth and seventeenth centuries their desperation relaxed, at the very time that interest rates fell. The last famines in England (Scotland was later) were in the 1590's (Peter Laslett, 1965, ch. 5). From the history of storage, in other words, one can infer that the interest rate had fallen quickly and deeply, an early stirring of modern economic growth.

I. The Direct Evidence that Storage was Small

The history of European storage has been neglected because the materials for its study appear so unpromising. One can learn a little from archaeological studies of grain storage, chiefly what the bins looked like, and a little from pretty tales and folklore. In 1540 at Nuremberg, for instance, Charles V tasted bread made from 118 year-old grain. It was proverbial in the fifteenth century, again, that "Winter alle etes/That summer begetes." Aside from such scraps there is little, and the subject has therefore been left to bald assertion and counterassertion. A leading student of the medieval French economy, Georges Duby, asserted flatly in 1962 that medieval Europe "did not know how to store grain or accumulate reserves" (p. 135). A leading student of the medieval Swedish economy, Eli Heckscher, asserted with equal confidence in 1941 that it was on the contrary a "storage economy" (p. 10). Neither provided evidence, nor emphasized the distinction between storage for consumption during the year after the harvest and storage for consumption in later years, the "carryover." The entire crop was, of course, stored for six months on average, because it needed to be eaten, which might justify calling any economy without continuous harvests a storage economy. Yet the carryover could be zero, with no reserves accumulated on the eve of the new harvest. It is carryover that is most to the point. Carryover smooths consumption, provides insurance, and links one year economically to the next.

---

*Professor of Economics and of History, University of Iowa, Iowa City, IA 52242, and Assistant Professor of Economics, Texas A&M University, College Station, TX and Economist, Federal Trade Commission, Washington, D.C. We thank Richard Ynteria of Chicago for excellent research assistance, and Alan Wiley and Stefano Fenoaltea for helpful comments. Earlier drafts were subjected to the friendly and productive scrutiny of seminars at the Canterbury meetings of the British Agricultural History Society, Australian National University, Berkeley, British Columbia, the California Institute of Technology, Chicago, Harvard, LaTrobe, the Law and Economics Center of Emory University, Maryland, Pittsburgh, Stanford, Texas A&M, Virginia Polytechnic Institute, and Washington. The research was supported by a grant from the Columbia University Center for the Study of Futures Markets.

The direct evidence on the size of the carryover is thin, although what there is suggests that it was small. The evidence must come from those who knew how to write and who wished to write down their doings, the lords and monks and burghers. The so-called "account roll" was an annual report to the lord's auditor by his bailiff combining elements of a balance sheet and an income statement, and in particular reporting fully on the disposition of the year's harvest of grain on the lord's own farm (his "demesne"). Accounts have been published for a dozen or so English estates out of the hundreds that exist in manuscript. The simplest way to use them is to look for the amount of "old grain" on hand at the reporting date, traditionally just after the harvest, and to divide that amount by the crop. The conclusion is that the carryover was small. For example, usable accounts of wheat from 170 of the years from 1208 to 1448 at Crawley, Hampshire contain only nine mentions of old wheat in any guise (N. S. B. Gras, 1930, pp. 339–43: "old," "from the previous year," "remain," "remain in sheaf," "left over"). These nine amount to something under 2 percent of the wheat produced in the 170 years in total.

An apparent difficulty with such evidence is that it is a lack of evidence. The rolls do not always say "old grain:nil." They often say nothing, leaving the observer to make the inference that nothing does mean nil. Yet William Beveridge was willing to make it: "Grain remaining from a previous year does not often occur, and if it does it is noted" (Beveridge, 1927, "Notes on Sources"). And more recent scholars agree: "bailiffs were appearing before auditors who had last year's account at hand. Had any corn been left on last year's account the bailiff would be required to account for what happened to it" (Eleanor Searle).

After a series of exceptionally good years, of course, the carryover might well be substantial. On the estates of the Bishop of Winchester in southern England, the years up to 1223 clearly were good ones. The manor of Wycombe carried over more quarters of grain ($246\frac{7}{8}$) than its entire crop (213), as did Ecchinswell and Burghclere (Beveridge papers, box 32, A49). The other manors carried over less, but a lot. The fifteen manors with usable accounts in 1223 had old grain of 1,561 quarters while harvesting 2,742 quarters, or carryover of fully 57 percent. But the usual case was no carryover at all. In 1220, only six out of seventeen manors had any carryover; in 1225, so soon after the remarkable prosperity of 1223, only four out of more than thirteen did; in 1236 only three out of eighteen. Not every year has usable accounts on the Winchester manors, but those that do most commonly have no single instance of carryover: 1226, 1231, 1232, 1235, 1244, and on into the century, punctuated by occasional bonanzas (such as the year 1256 at the big manor of Taunton: 624 quarters of grain "of the second year" and even 182 "of the third," as against 824 quarters fresh from the fields). The bonanzas were rare, and not to be relied upon.

These, of course, were bonanzas to the rich, not to the average man. One would like to know how if at all the other 95 percent of the population carried over grain from one harvest to the next, but the evidence is even thinner than that for the larger owners of land. Certainly there was no sharing out of what the richer folk might hold. Around 1500, a priest scolded a rich and selfish ploughman (this at Lent, incidentally, long before the next harvest):

> Thou knowest that of corn is great scarceness,
> Whereby many for hunger die, doubtless,
> Because they lack their daily bread—
> Hundreds this year I have seen dead;
> And thou hast great plenty of wheat
> Which men for money now cannot get.
> [Celia Sisam and Kenneth Sisam, p. 516]

The priest proceeds to swindle the ploughman out of his plenty, for the benefit of the poor. But the poet is silent on the size of carryover.

Stefano Fenoaltea (1970) has attempted to use other sources (though still pertaining to the rich) to estimate the carryover, namely, the number and dimensions of monastic barns. He estimated the carryover to be very high, concluding that "the monastic barns alone could hold enough grain to feed England's human population for over a year and a half" (p. 139). His reasoning is in error, for it does not allow for the storage of seed, which was at medieval yields fully a quarter of the crop. The correct arithmetic is as follows. Suppose that the barn's capacity was commonly exhausted.[1] Suppose too that carryover was some fraction, c, of the year's whole consumption (taken to be 3.0 units, with 1.0 for seed, the whole output being 4.0). If the capacity of the barns was as much as 1.5 times annual consumption, the accounting of the crop is

$$\text{Barn Capacity} = \text{Consumption} + \text{Seed} + \text{Carryover}$$

or,

$$(1.5)(3) = 3 + 1 + c3$$

The carryover share c implied by this capacity of one-and-one-half times consumption is 1/6, not 1/2.

Since carryover is calculated here as a residual it is sensitive to small errors, and the errors in estimating capacity are not small (for instance, grain was usually not threshed before being stored, but sometimes was). At a capacity of 1.33 rather than 1.50 times consumption the arithmetic implies there would be no room for carryover at all; at a capacity of 1.66 times consumption the carryover fraction would be .33. That is, the fraction varies from zero to .33 when the capacity estimate varies plus or minus by a mere one-quarter or one-fifth.

The decisive objection to a large carryover, however, is evidence on the frequency of starvation. The distribution of production is known to be normal with a standard deviation of 35 relative to an average of 100; starvation is known to have occurred at a consumption of 50 relative to 100 (McCloskey, pp. 141–45). One can use tables of the normal distribution to reckon the waiting time to starvation, seeing whether large carryovers give reasonable waiting times. For example, if the carryover against crop failure was 10 compared with an average consumption of 100, then a peasant could survive any single year (i.e., consume 50) with a crop of 40 or above. A crop below 40 would cause him to starve, as would two crops in succession of, say, 44 and 44 (since the carryover, depleted by 6, could not be rebuilt before disaster hit again). But large carryovers do not give reasonable waiting times: see Table 1.

The actual waiting time calculated from the Anglo-Saxon Chronicle and other sources is on the order of 10 or 15 years (see McCloskey, p. 144). The 30 to 476 years of waiting time characteristic of high carryovers

---

[1] The capacity of barns would not in fact have been exhausted, since barns were built for the *maximum maximorum*, the peak harvest year as well as the peak month within a year. Transport costs were high, implying that it would be optimal to overbuild local capacity. Almost never would there be a year using every barn in England to capacity, since each region would not have the same bumper crop. More commonly, down to early modern times in much of Europe, capacity in one region might be fully utilized while it went abegging, literally, in another.

TABLE 1—EXPECTED NUMBER OF YEARS BETWEEN FAMINES FOR VARIOUS CHOICES OF CARRYOVER AND THE FAMINE LINE
(Average consumption = 100)

Famine Line relative to 100	Carryover (Percent) 5	10	20	50
50	22	32	44	476
60	13	16	28	28+

are too long, sometimes absurdly long. Even a starvation level of 60 does not easily reconcile high carryovers with the short waits observed. That the medieval economy was haunted by starvation, in short, implies that it had little in its stores of grain.

Particular episodes of famine suggest that carryovers were at most on the order of 5 percent of consumption, not 50 percent, surely, or even 20 percent. The years leading up to the autumn of 1315, when the crop was very poor all over Europe, were not bad. Yet Henry Lucas notes that by the spring of 1316 "the old stocks became completely exhausted" and wheat sold in England for 40 shillings a quarter (compared with 5 shillings in a normal year) (1962, p. 55). He gives a table of deaths in Ypres, rising sharply in the spring. Where was the six-month buffer stock? On the demesnes of the Bishop of Winchester, net yields in the bad years 1315 and 1316 were around 45 percent of the 1314 yield. A buffer stock as large as six months of consumption would have easily permitted consumption in these 2 years at well above the starvation level (50 percent of the average consumption). Yet 1315 and 1316 was burned in men's memories, the worst famine in European history.

II. Using Prices to Show that Storage was Small

The notion of a "storage economy," then, is not favored by the usual evidence, though the usual evidence is scanty. Another class of evidence bearing on the issue, however, is available in large and elastic supply: prices of grain. For no time after the twelfth century is it difficult to get the prices at which manorial farms, monasteries, Oxford colleges, the King's household, chartered towns, and other English institutions bought and sold wheat, barley, oats, and rye. From Poland to Portugal, indeed, medieval records yield prices in limitless array, dated, published, affixed to goods of ascertainable quality, and unused for historical purposes beyond the measurement of long-term inflation. The outmoded yet strangely resilient notion that the Middle Ages were ages of "natural economy" unused to trade, money, and prices looks odd beside such an outpouring. Here in quantities beyond the wildest dreams of intellectual avarice are the statistics of a commercial civilization.

One simple way of using prices is to ask whether the amount they fluctuated from harvest to harvest was consistent with a large and therefore price-damping carryover. They fluctuated a lot, which is inconsistent with it. Annual coefficients of variation calculated from the essentially trendless series of prices of wheat at two manors in Hampshire from 1245 to 1350 for 20-year periods ranged from .20 to .43 (J. Z. Titow, 1969, pp. 97–99). The coefficient of variation in wheat prices at Philadelphia, 1800–25, was .26, and at New York in roughly 20-year periods from 1825 to 1914 ranged from .16 to .34 (U.S. Bureau of the Census, 1975, series E123). The typical fall from medieval to modern times, then, was from about .30 to about .24: a variety of climates newly accessible by cheap transport (the Baltic, for example, in early modern times) was a substitute for a large carryover and had the same effect. The only oddity is that the effect was not even larger.

A further step along the same line of reasoning tests for the influence of carryover by regression methods. If carryovers were unusual, then in a very good year one would expect carryover stocks to be built up, depressing prices in the next crop year below what they would be had the previous crop not been good. One would expect a routine of large carryovers, on the other hand, to allow little impact on prices of a good crop. For a sample of 39 years in the thirteenth and fourteenth century in southern England, one's expectations are fulfilled, if one expected small carryovers:

$$\ln P_t = \underset{(0.28)}{3.59} - \underset{(0.13)}{1.15}\ln Q_t - \underset{(0.11)}{0.205}\ln Q_{t-1} - \underset{(.12)}{0.301}\ln Q^*_{t-1}.$$

The fitted equation (standard errors in parentheses; $R^2 = .76$; Cochrane-Orcutt applied) says that price now (P_t) is reduced very strongly by a high yield now (Q_t), much less strongly (though definitely) by a high yield last year (Q_{t-1}) and strongly by yields last year more than one standard deviation above the mean (Q^*_{t-1}). In other words, last year's yield can be divided into two different variables, yields more and less than one standard deviation above the mean. The equation permits a kink, testing whether unusually high yields lead to unusually high carryovers and unusually great depression of next year's prices. They do. In short, carryovers existed, doubtless. But they were small and sporadic, not sustained at such high levels as 30 or 50 percent of consumption. They were probably closer to 5 percent or less, except after a spectacularly good year.

III. The Cost of Storage was High

The question is, why?—why was carryover so small? The answer is that it was expensive, which fact in its turn buttresses the shaky calculations of its smallness. As has been noted, the cost of storing a bushel of wheat is the cost of the barn per bushel plus the percentage rotting in storage plus the expected percentage loss of capital value due to falls in the price per bushel plus the opportunity cost of the interest forgone on the sum expended on the bushel. In what follows the equation will be broken down into the parts, examining what little direct evidence there is for each. At present it is necessary to accept only that the total carrying cost must be earned in equilibrium. If the storers of grain do not make systematic errors in predicting what prices will be—that is, if they are in the fashionable parlance "rational"—then the price of grain will in fact march up at the monthly carrying cost. It can behave on average no other way. If, on the contrary, storing a bushel for four months from September to January were to earn persistently less than four times the monthly cost (suitably compounded), then storers would store less, driving the September price down and the January price up until the equality was reestablished. The harvest, or the signs of the harvest, break the logic, which begins anew at a price suitable to the size of the new crop. So it goes year upon year, in sawtooth fashion. The slope of the teeth is the cost of storage.

All that is needed is the average slope of prices within the harvest year. The account rolls again are ultimately the chief medieval source. They usually give prices without date ("of wheat sold, whereof 2 quarters the price per quarter 6s. 4d., 44 quarters 4½ bushels the price per quarter 6s 8d" and the like). But sometimes the prices are dated, commonly by saint's days. The largest collection of raw, dated prices is J. E. Thorold Rogers' *The History of Agriculture and Prices in England*, published in 1866, and long recognized as a rich and reliable mine for data. Volume I contains about 170 pages devoted to grain prices in the late thirteenth and fourteenth centuries, or some 27,000 quotations. Some 8 percent of these, about 2,000, have more or less precise dates attached, about fourteen dated quotations on average for each year 1260–1400.

The facts for wheat are most voluminous, because wheat was the commodity of commerce. The monthly rates of change of prices can be calculated from comparisons over various pairs of months. There are 1,075 such pairs observed in the same village and the same year. For instance, one pair is the rise from May to July 1331 in the price of wheat from 6s. 8d. to 6s. 10d. per quarter of 8 bushels in Elham, Kent. There are a total of 22 cases of May to July comparisons in the Rogers data, the first in the crop year beginning in 1272 and the last in 1356, with an average ratio of 1.055 (the standard deviation around the average is high, .160; the standard error is .034). Two-thirds are from Cuxham, Oxfordshire (an Oxford college owned the village); the rest are scattered about southern England. The average

TABLE 2—ENGLISH WHEAT PRICES IN THE THIRTEENTH AND FOURTEENTH CENTURIES, PERCENTAGE RATES OF CHANGE PER MONTH

From	To October	November	December	January	February	March	April	May	June	July	August
September	7.5	8.5	2.8	2.3	1.6	2.0	3.4	4.8	2.0	2.4	1.7
October		3.8	1.4	0.1	0.3	−0.4	1.6	1.6	0.6	3.9	0.6
November			6.4	2.2	1.5	1.2	1.6	0.8	0.7	1.5	1.4
December				6.3	6.8	2.8	2.6	2.6	2.8	0.7	0.9
January					1.9	3.2	4.0	1.4	1.5	0.1	0.8
February						3.7	4.1	4.0	2.4	2.0	2.4
March							3.6	2.2	2.1	2.6	2.8
April								1.2	0	2.8	−0.4
May									1.5	2.7	−1.5
June										4.2	4.0
July											1.5

Source: Authors' calculations, from 1,075 pairs of prices, 1260–1400.

TABLE 3—UNWEIGHTED AVERAGES BY MONTH OF ALL RATES OF CHANGE OF WHEAT PRICES, 1260–1400, ENCOMPASSING A PARTICULAR PAIR OF SUCCESSIVE MONTHS

September-October	3.55	January-February	1.91	May-June	1.52
October-November	2.25	February-March	2.04	June-July	1.78
November-December	1.85	March-April	2.12	July-August	1.29
December-January	2.03	April-May	1.78		

Source: See Table 2.

monthly rate of change of prices from May through June to July was therefore the solution, r, of $e^{2r} = 1.055$, or $r = 2.70$ percent per month, some 38 percent per year.

One way of displaying the results is in a matrix of pairs of months, as in Table 2. One way of summarizing the 66 averages is to say that their average in turn is 2.37 percent per month (32 percent per year), though with a high standard deviation (1.87 percent, a coefficient of variation of .79 around the average) and high standard error (0.23).

Another way to summarize the averages is to extract all the information in the matrix concerning a rate of change between a particular pair of consecutive months, to detect any pattern of seasons. The entries relevant to the March-April comparison, for instance, are all those northeast of the March-April entry, 3.6. Adding them together without weighting double-counts many times over and gives the same weight to an entry such as November-August, which merely passes through the March-April comparison, as to one such as March-May, which is closer to the nub of the issue. Without theoretical light it is hard to see the merit of one average over another. For what this one is worth, the result is given in Table 3. The storage costs are high at the outset, falling to around 2 percent per month (about 30 percent a year). That the cost is lower in the spring probably reflects the arrival of definite news about the next harvest. The lowering may indicate the size of the risk premium required earlier in the year, when news is scant and long-term averages are the only guide to how prices will change.

Averages by length of the gap between months, reported in Table 4, seem also to be telling. What exactly they tell is not altogether clear. Very likely the fall as the comparisons get longer reflects the fixed cost of putting grain in storage. Possibly, too, it represents a selection bias in the evidence, since a more rapid rise in prices over a short period would cause more transactions to be undertaken by the manors and other institutions involved, and a greater likelihood that a pair of prices would fall in successive

TABLE 4—AVERAGES BY LENGTH OF GAP BETWEEN MONTHS COMPARED

Number of Months Gap[a]	Monthly Rate of change[b]	Number of Observations	
		Actual	Expected
1	3.77	254	179
2	3.57	204	163
3	2.06	166	146
4	1.58	138	130
5	1.21	99	114
6	1.61	69	98
7	1.51	64	81
8	2.06	44	65
9	2.47	20	49
10	1.55	13	32
11	1.73	4	16

Source: Authors' calculations of wheat prices.

[a] For example, January–June = 5.

[b] Shown in percent.

months. The form of the matrix of comparisons implies the last column, the expected number of observations at each number of months gap. For example, the 11 elements on the diagonal are one-month gaps, 11 out of 66 total. That is, 11/66 of the 1,075 observations, or 179, should be one-month comparisons under the null hypothesis of equiprobability of each element. There is a marked overrepresentation of brief gaps (close to the diagonal). One could examine the runs of prices that are producing the brief gaps to see if it is plausible that they are themselves a consequence of rapid price rises. In any event, the nature of the sampling biases in the Rogers data are worth pursuing: weighted by the number of observations the average rise is a 2.55 percent per month (35 percent a year), while the average for pairs more than two months apart is only 1.75 percent (only 23 percent per year).

The tentative conclusion from even these crude methods is nonetheless plain. The whole cost of wheat storage was from about 2 to 3 percent per month, or from 27 to 43 percent per year, closer perhaps to 30 than to 40 percent.

The result is confirmed by a more elegant if less rich method, namely, a regression of prices from the same harvest year and place against the distance of days between them. It comes in two forms, depending on the placing of the constant term. Either

$$P_t = P_0 e^{rt} + c' \quad \text{or} \quad P_t = P_0 e^{c+rt}.$$

In other words, the price on the second date (P_t) is raised either by a constant cost per bushel of putting wheat in storage (c') or by a constant cost per shilling value (c). The method of estimation of the two forms is different, but their results do not differ radically in the central parameter, r, the rate of growth of prices (here, per day). The cost-per-bushel form required a nonlinear method, giving

$$P_t = P_0 \exp(.001165t).$$

The rate of change is 0.1165 percent per day, or $(1.001165)^{365} = 1.53$, or 53 percent per year. The constant-per-shilling form required logarithms and fitting by least squares, giving

$$\ln(P_t/P_0) = .0082 + .00083t.$$

The R^2 is very low (.07), unsurprisingly so. The standard error of the coefficient on t (that is, time in days) is one-half the value of the coefficient. The implied rate of change is $365(.00083) = .30$, or 30 percent a year. The econometrics echoes the ambiguity of the earlier calculations: the cost of storage is probably around or above 30 percent a year, but where exactly around or above depends on exactly how one wishes to look at the facts.

The regression results are based on 61 observations (as against 1,075 in the monthly statistics), about one-half of the exploitable pairs of dated prices in the same location and year. Little would be gained that is not captured in the monthly averages by adding observations dated simply "March" or "Spring." Were they nonetheless added, a definite downward bias from errors in variables would follow (although it is clearly no trick to estimate the error variance from saying "March" when some definite date in March is meant). Four observations of extreme drops rather than rises in prices were junked as outliers: including them turns the

TABLE 5—AVERAGE RATES OF MONTHLY GROWTH IN PRICE FOR MINOR GRAINS[a]

	Percent Rate of Change[b]	Standard Deviation	Number of Pairs of Months with Data[c]	Definition or Use
Rye	4.17 (1.31)	6.16	22	A bread grain inferior to wheat; black bread
Mixtil or Maslin	3.26 (1.36)	6.64	24	Rye and wheat sown together
Drage or Dredge	4.64 (1.51)	7.38	24	Oats and barley sown together
Malt	5.8 (2.1)	3.7	3	Sprouted barley
Drage Malt	1.		1	
Beans	2.5 (3.3)	8.8	7	Pod-bearing vines
Vetches	8.3 (3.1)	7.5	6	Mainly for animal feed
Peas	3.8 (5.)	16.	10	

Source: Authors' calculations.

[a] Thirteenth and fourteenth centuries, September through June only; excluding the famine years 1315–16.

[b] Standard errors are shown in parentheses.

[c] Except for rye, nearly equal to all the quotations.

equations into rubbish, in that the standard errors shoot up and the cost of storage becomes negative. Finally, the procedure truncated the year at 240 days after the harvest was finished (notionally, September 1), because including May, June, July, and August adds, again, more noise than music. The fact is itself significant, helping to confirm the model being used. From May onward, it would become increasingly clear how large the harvest was in fact going to be, causing prices to fan out to their ultimate destinations. No longer would the average experience of past years govern the rate of rise of prices.

Wheat was not the only grain. Barley, for example, was grown to the same or greater extent as wheat (the word "barn" derives in Anglo-Saxon from "barley"). It was used commonly for bread, occasionally for animal feed, or malted and brewed into the quantities of weak beer that made up a large part of the medieval diet. The analysis of barley gives evidence of very large price rises. The one-month comparisons beginning in September, October, November, and June, for example, average a 10 percent rise per month. The minor grains—minor at least as marketed goods—tell a similar story. A rise in round figures of 4.0 percent per month for the minor grains appears reasonable, that is, 60 percent per year.

The results for oats (usually a bread grain) are more full, and tend to the same conclusion that the rise was very high. The average of the 42 monthly comparisons (namely, all except those ending in July and August, when the harvest began, and three months with no information) is a 5.77 percent monthly rate of rise, with a standard deviation of 4.52 and a standard error of the mean of 0.70. A 95 percent confidence interval for oats would run from 3.1 percent to 5.9 percent per month (from 45 to 99 percent per year).

That other grains have so much higher costs of storage than wheat requires explanation. It is possible that lesser grains did not store as well as wheat, or that the dealers in wheat were better placed in the capital market. The cost of storage, in other words, might actually have been higher. But there is a simpler explanation. Because the other grains had to compete with wheat for storage space, and because the space was paid for by the price rise, the absolute price rise per bushel for the other grains would have to have been the same for wheat. Wheat sold for twice as much per bushel as did other grains. In percentage terms, consequently, the monthly rise would have to be higher for the lower priced grains. And wheat stored more compactly than barley. In any event, the annual cost of storage of grain was at least 30 percent per year, and could well have been higher.

IV. The Cost was High Because the Interest Rate was High

The cost of storage fell in modern times. One can duplicate the calculations for the sixteenth century and after, the data being of course plentiful. In the Oxford town market, for example, there are monthly figures from 1618 to 1644. November was the low month for wheat prices, August the high, with a rise between them of only 8.4 percent (the standard error is 3.7), or only 0.9 percent per month and 11.3 percent per year. In Namur, Belgium, 1614–92, the August to June rise for wheat is 12 percent, or only 14 percent a year. In Diest, in the Brabant province of Belgium, the rates of annual rise by crop, 1718–36, are wheat 10.9 percent, barley 14.8, rye 14.5, oats 15.4, and buckwheat 18.1. The instances could be multiplied indefinitely, with the same outcome. The cost of storage, plainly, was much lower in the seventeenth and eighteenth centuries than it had been four centuries earlier.

To explain why it fell, one must turn to the components of the cost—the percent spoilage, the barn costs, and, above all, the prevailing rate of interest. Direct scrutiny of these confirms the impression of very high costs in the Middle Ages and suggests why storage eventually became cheap.

What, then, accounts for the 30 percent cost per year or more? The rotting of grain was probably not more than 10 percent per year. This at any rate is the figure suggested by the scraps of evidence nowadays on poor countries, whose techniques of storage are little better than medieval. The latest estimate comes from a study in 1978 by the National Academy of Sciences:

> Experts involved in the preparation of this report resisted extrapolating post-harvest loss estimates to national or global levels because general estimates cannot be supported with statistically significant data. For planning purposes, however, 10 percent is cited as an average minimum overall loss figure for cereal grains and legumens. [p. 8]

The figure is the loss from harvest to consumption, implying that it would be on average the cost for one-half year (half a year being the average time a bushel of wheat spends in storage if there is one crop a year). The annual cost, which is what is relevant here, would therefore be higher. And the 10 percent is an "average *minimum*" itself, implying again that the average average is higher. On the other hand, these are losses for tropical countries. They would be smaller in temperate countries—and indeed the same report gives 10 percent as an average (not a minimum) for Rhodesia, ranging up to 52 percent on occasion in hotter places such as India (p. 85).

No changes occurred in the character and cost of barns that would explain the fall in total costs from medieval to early modern times. Even in the sixteenth and early seventeenth centuries, Peter Bowden argues, "adequate storage facilities...were lacking," by which he means "adequate" facilities of modern type (1967, p. 816). M. W. Barley emphasizes a continuity of rural building types extending well into the late sixteenth century. It was houses, not barns, that evinced the prosperity of Elizabethan landlords. Writing of rich yeomen in East Anglia even in the early seventeenth century, he asserts that "while many such men enlarged or modernized their medieval houses, the barns they inherited were often perfectly adequate" (1967, p. 744). In the open field areas of the Midlands and the North, there were still in the seventeenth century "hovels" or "helms" or "belfries," that is, granaries on posts with adjustable roofs: but they had been "a regular feature of Germanic villages in the migration period and later; the adjustable roof is shown in manuscripts of the later Middle Ages" (Barley, p. 744). The timber (and labor) with which one might make a barn fell in price relative to grain by about 50 percent from 1450 to 1650 (Bowden's figures, p. 862), but it would be surprising if the barn costs—except perhaps on account of the high interest rate on the investment in structure—were much of a factor in storage costs to begin with.

The direct testimony on the rate of interest in the Middle Ages is not easy to interpret. The traditional usury limit in England for what amounted to pawnshop loans on good security was 2 pence per pound per week, or 43½ percent per year at simple interest and 54 percent compounded. The figure is reproduced in all discussions of medieval interest rates. So too is the history of the law of usury and its evasion. The prohibition of usury was irrelevant: that the sin of taking interest should be committed frequently is no more surprising than that the sin of adultery was. Interest rates far above 50 or 100 percent were common on personal loans, though the rates on commercial loans among the rich were at more modern levels. It is this division of the market that makes the evidence cloudy. In poor countries, seven centuries later very high interest rates by moneylenders coexisted with subsidized loans at 5 or 10 percent from government agencies (U Tun Wai, 1956, 1957). What seems clear about the European experience is that, at least in the portion of the market in which ordinary peasant cultivators found themselves, the early modern period witnessed a great fall of interest rates. By the late seventeenth century in Lincolnshire, for example, probate inventories attest that "nearly everyone with surplus cash appears to have let out at interest," at only 4½ to 6 percent per year, when rates in London were 5 percent (B. A. Holderness, 1975, pp. 108, 97). At the *haute finance* end of the scale, there is a halving of the interest rate in Northern Europe from the thirteenth to sixteenth centuries, although the clarity of the fact is muddied by the vagaries of war and the varieties of financial instruments (Sidney Homer, 1977, p. 142).

The evidence from grain prices is in important senses better than that on financial instruments. It is uniform over many centuries. It is available for all parts of Europe. And it is free of default: one cannot default on one's own investment in stored grain, or in whatever investment alternative to stored grain was the opportunity cost of investing in grain. Such an investment is not, of course, free of risk, for stored grain faces its own hazards. One was taxes, which were assessed (usually in the fall) on any asset the taxgatherer could spot. Another was the risk of a fall in prices, since prices did not march up mechanically. More subtle calculation might extract the premium for the price risk from the data, though such subtleties are hard to take seriously. The variability of grain prices, as was noted earlier, did fall a little, though hardly enough to account for a revolution in the risk premium demanded. A measure of its size would have other uses, in suggesting how medieval people viewed the hazards of life more generally. No matter: even the pure interest rate in the Middle Ages will be found to have been very high.

V. Other Assets, such as Livestock, Reveal a High and Falling Interest Rate

The pricing of any asset embodies the interest rate. Statistics on livestock, for intance, are rich sources of indirect evidence, supporting the conclusion that the medieval interest rate was high and fell sharply by the sixteenth century. The statistics have survived because livestock were worth keeping records on—easy to steal, often put in the care of others, and frequently bought and sold. The data include herd and flock size and composition by sex and age, death rates, slaughter rates, and prices of inputs and outputs (hay, grazing land, straw; labor; milk, cheese, butter; wool, hides). All manner of ratios among these things are governed by the prevailing rate of interest, and can be forced to tell what it was.

One ratio, for example, is of net revenues to the value of the capital invested in animals. It is, of course, the interest rate. On the Crawley, Hampshire estate of the Bishop of Winchester in 26 years, with the requisite facts from 1208–09 to 1254–55, the typical yearly stock of sheep was worth about £56 (on the order of 22 man-years worth). The earnings attributable to the sheep on average each year were £4.2 from animal sales, £14 from wool, £1.5 from pelts, and £5.5 from sheep's cheese, or £25.2 in total. Since Crawley was a market-oriented place devoted largely to sheep raising, it is unlikely that much product was kept unsold for home use. The flock grew its own replacements, which sidesteps the problem of depreciation. The one significant hole in the calculation is the yearly cost of keeping the animals. Sheep are the best case if one is forced to guess, because unlike cattle they do not require intense husbandry animal-by-animal and, unlike horses, they do not (except lambs) require feed grain. The interest rate on this reckoning is 45 percent (or a little under if allowance is made for labor, grazing, and fencing costs).

A calculation that skirts the difficulty of knowing costs of upkeep asks how much interest rates fell. The claim is that the sixteenth century witnessed a great fall in interest rates. Bowden's indexes of prices of sheepish things (see p. 848ff) rise during the great inflation of the Long Sixteenth Century from 1450–99 ($=100$) to 1640–49 thus: sheep themselves, the capital good, to 681; the revenues from using the capital good, wool and sheepskins, to 396 and 372; the costs of using it, hay and straw, to 768 and 612. Note that revenue fell relative to costs. Consider that the interest earnings on an investment in one sheep are equal to annual revenue minus cost of upkeep: $iP = R - C$. By advanced algebra it follows that $i = R/P - C/P$. The change in interest, i, must equal the change of the ratios on the right. Wool and sheepskins are revenue, and fall to one-half their former value relative to the price of sheep. Hay and straw are costs, and stay the same relative to sheep (never mind labor, which falls even sharper than wool but is not a large input). If C were known, as is R and P, there would be no trick to finding i. It is not. But knowing the terminal i and the initial R is enough. Suppose (as would be easy to show) that, by 1640–49, the interest rate was, say, 15 percent on such an investment as one sheep. Take the price of that sheep to be the numeraire (thus its price is always 1.00), take the annual revenues before the fall in interest rates to be .45 relative to the price of 1.00 (recall that this was in fact the early thirteenth century ratio at Crawley), and take the fall in revenues to be a fall to one-half this level (as the price indexes imply). Thus in 1640–49, after the fall,

$$.15 = .225/1.0 - C/1.0,$$

where the .225 is half of the original .45. The value of C after the sixteenth century must be .075. The price indexes imply that it did not change relative to the price of sheep. Before the fall of relative revenues, then, the interest earnings were $2(.225) - .075 = 37.5$ percent, as against 15 percent after. The interest rate, as usual, is well above 30 percent.

This conjuring trick would warrant no great confidence standing alone. It does, however, serve to confirm the high net interest found by other methods (and, incidentally, the low costs of upkeep C per sheep relative to the medieval values of R). A more sturdy statistic, and another variation on the theme that cattle are capital more than merely etymologically, is the ratio of the hire rate to the price. A milk cow, in particular, sold for 10 shillings or so in the thirteenth and fourteenth centuries, while renting on various terms for anything from 4 to $6\frac{2}{3}$ shillings (Rogers, 1866, pp. 361, 397; 1884, p. 94, and Dorothea Oschinsky, p. 427). Roughly speaking, the figures imply interest rates from 40 to 67 percent per year. But dividing rent by price is too crude. For one thing, the owner paid for the feed (see Rogers, p. 94), so as usual the calculation runs afoul of the cost of upkeep. For another, cows were not immortal. The second point can be handled by cutting off the stream of discounted returns at the 4- or 5-year average working life of a cow. The price-rental ratio for an asset with a life of n years at an interest rate of i is $(1/i)[1 - 1/(1+i)^n]$. A price-rental ratio of 2 (price of 10s. and rental of 5s., not allowing for deductions for feed) and a life of 4 years implies an interest rate of 35 percent.

Other approaches are possible, though not pursued here. Calves grow up to be "hoggasters" and *bovetti* (heifers), and thence cows and oxen: the project of letting them grow must earn the rate of interest. The death of mature cattle and sheep is not usually an accident (though it should be remembered that cattle were raised chiefly to pull plows, not to fill plates): the optimal slaughter date also depends on the rate of interest. Plowteams were rented: the rental-price ratio reflects the rate of interest. Land itself was both rented and sold: being a perpetuity, the rental-price ratio is the rate of interest. (By the seventeenth century, in France land often rented for 5 percent of its sale price.) It is to be expected that all these will show a remarkably high interest rate in the thirteenth century and a low rate by the sixteenth century.

VI. The Significance of High Medieval Interest Rates

The hero—or villain—of the tale, in short, is the rate of interest. A fall of interest rates from, say, something over 20 percent to something under 10 or 12 percent must have been significant for the history of Europe and for the onset of modern economic growth. An economist can easily believe this. But it is unclear exactly how it was significant. If interest rates were so high, why were Europeans so poor? A few years of moderate abstention from consumption invested at over 20 percent would make a man rich. True, men of all sorts, bond and free, did in fact scale the social ladder quickly in the Middle Ages. The economy and society are known now to have been much more fluid than was once thought. But the point is that the whole society could and would have done so. Perhaps as is commonly argued nowadays in poor countries the menu of opportunities dropped off sharply in quality after the first few items. Low savings rates might explain it, too, though the lack of savings in medieval times is easy to misapprehend. Peasants worked hard at ploughing fallow land and cleaning ditches in aid of yields quite far removed, and this was saving. Since the yield to seed ratios in the Middle Ages was only 4 or so grains returned from each planted, a quarter of the yield had to go back into the ground as investment in seed. On both counts (food being most of income), the saving entailed by medieval agriculture was enormous, albeit brief and direct. Medieval people saved a lot.

A more conclusive line of argument among several possible is that high interest rates made it difficult for peasants to depend on the capital market at all. The reasoning is that loans would be sought before the harvest and paid back after. In the meantime the price of grain would fall, by some 30 percent in a normal year, and much more after the very years of dearth in which most loans would be made. Money that bought little grain before the harvest would have to be earned back by the borrower after the harvest by selling the grain just grown at prices that bought little money. John Waryn of Oxford borrowing 10 shilling on February 1, 1322 to buy 8 bushels of rye might pay back the debt six months later at a monthly rate of 2 percent in money, the rate earned on stored grain. But the price of rye per bushel might have fallen after the harvest from 10 shillings to 7 shillings for 8 bushels. In fact it usually did. To pay back the 10 $(1.02)^6 = 11.26$ shillings on August 1, he would have to have given up 11.26 $(8/7) = 12.87$ bushels of his new crop. For the 8 bushels on January 1, he would pay 12.87 bushels on July 1. The monthly interest rate is not the monetary 2 percent (27 percent per year) but 8.2 percent a month, an unpayable 260 percent per year. If the price fall after the harvest were smaller, say the 12 percent that becomes typical in early modern times, then the real interest rate is 56 percent per year, no small rate but at least potentially payable. The effect of improved storage (itself an effect of the fall in money interest rates) was to cut radically the commodity rate of interest. Before the cut, the peasant was forced back on expensive methods of self-insurance such as scattering of strips in order to avoid having to borrow. All manner of medieval institution —even serfdom itself—can be viewed as replacements for an expensive capital market. Loans among peasants were by no means unknown in medieval times, especially late medieval times (Elaine Clark, 1981). But, by the sixteenth century, they were common, the custom of scattering strips was beginning its long decline, and other medieval institutions were being cleared away.

There is one more point about the relations of medieval villagers revealed by their interest rates. If interest rates in terms of money were at the levels reported here, and higher in terms of bread, the brotherly image of the medieval countryside once taught in school and still taught in Hollywood becomes cloudy. Where is Jack's brother peasant when Jack must borrow from the moneylender at 27 percent in money and well over 100 percent in real terms to survive? The point is one more among many against the myth of the medieval community, that charming product of nineteenth-century German scholarship and speculation on the Middle Ages. The myth of the *Markgenosse*, a putative community of free Germanic citizens, and all that is alleged to follow from it has been under attack for three-quarters of a century. By now most medievalists agree with David Herlihy that "research has all but wiped from the ledgers the supposed gulf, once considered fundamental, between a medieval manorial economy and the capitalism of the modern period" (1971, pp. 154–55). So too the supposed gulf in attitudes towards one's neighbors. "You know, Ernest, the very poor in the Middle Ages were different from you and me." "Yes, Scotty, they were poorer," because of the interest on corn.

REFERENCES

Barley, M. W., "Rural Housing in England," in J. Thirsk, ed., *The Agrarian History of England and Wales*, Vol. IV, *1500–1640*, Cambridge: Cambridge University Press, 1967, 696–766.

Beveridge, W. H., "The Yield and Price of Corn in the Middle Ages," *Economic History*, May 1927, *1*, 155–67.

______, Papers of the Project on Price History, London School of Economics, 1925–30.

Bowden, Peter, "Statistical Appendix, A: Prices of Agricultural Commodities," in J. Thirsk, *The Agrarian History of England and Wales*, Vol. IV, *1500–1640*, Cambridge: Cambridge University Press, 1967, 814–55.

Clark, Elaine, "Debt Litigation in a Late Mediaeval Village," in J. A. Raftis, ed., *Pathways to Medieval Peasants*, Toronto: Pontifical Institute of Mediaeval Studies, 1981, Papers in Mediaeval Studies 2, 247–82.

Duby, Georges, *Rural Economy and Country Life in the Medieval West*, Columbia: University of South Carolina Press, 1968 (originally published in French, 1962).

Fenoaltea, Stefano, "Risk, Transaction Cost, and the Organization of Medieval Agriculture," *Explorations in Economic History*, April 1970, *13*, 129–51.

Gras, N. S. B., *The Economic and Social History of an English Village*, Cambridge: Harvard University Press, 1930.

Heckscher, Eli F., *An Economic History of Sweden*, Cambridge: Harvard University Press, 1963 (originally published in Swedish, 1941).

Herlihy, David, "The Economy of Traditional Europe," *Journal of Economic History*, March 1971, *31*, 153–64.

Holderness, B. A., "Credit in a Rural Community, 1660–1800," *Midland History*, No. 2, 1975, *3*, 94–115.

Homer, Sidney, *A History of Interest Rates*, 2d ed., New Brunswick: Rutgers University Press, 1977.

Laslett, Peter, *The World We have Lost: England Before the Industrial Age*, New York: Charles Scribner's, 1965.

Lucas, Henry S., "The Great European Famine of 1315, 1316, and 1317," *Speculum*, 1930, reprinted in E. M. Carus-Wilson, ed., *Essays in Economic History*, Vol. II, London: Edward Arnold, 1962, 49–72.

McCloskey, Donald N., "English Open Fields as Behavior Towards Risk," in P. Uselding, ed., *Research in Economic History*, Vol. 1, Greenwich: JAI Press, 1976, 124–70.

Oschinsky, Dorothea, *Walter of Henley and Other Treatises on Estate Management and Accounting*, Oxford: Clarendon, 1971.

Rogers, J. E. Thorold, *The History of Agriculture and Prices in England*, Vol. I, London, 1866.

______, *Six Centuries of Work and Wages*, London, 1884.

Schultz, Theodore W., *Transforming Traditional Agriculture*, New Haven: Yale University Press, 1964.

Searle, Eleanor, personal correspondence, 1982.

Sisam, Celia and Sisam, Kenneth, *The Oxford Book of Medieval English Verse*, Oxford: Clarendon, 1970.

Thirsk, Joan, *The Agrarian History of England and Wales*, Vol. IV, *1500–1640*, *Cambridge: Cambridge University Press*, 1967.

Titow, J. Z., *English Rural Society 1200–1350*, London: Allen and Unwin, 1969.

Tun Wai, U, "Interest Rates in the Organized Money Markets of Underdeveloped Countries," *International Monetary Fund Staff Papers*, August 1956, *5*, 249–78.

______, "Interest Rates Outside the Organized Money Markets of Undeveloped Countries," *International Monetary Fund Staff Papers*, November 1957, *6*, 80–142.

National Academy of Sciences, Board on Science and Technology for International Development, *Postharvest Food Loss*, Washington, 1978.

U.S. Bureau of the Census, *Historical Statistics of the United States, Colonial Times to 1970, Bicentennial Edition*, Part 2, Washington: USGPO, 1975.

Robert Giffen and the Irish Potato

By GERALD P. DWYER, JR. AND COTTON M. LINDSAY*

A fallacious consensus seems to be emerging on the subject of Giffen goods. The potato famine experienced in Ireland in 1845–49 is increasingly being cited as a likely instance of the occurrence of this logical exception to Alfred Marshall's "Law of Demand."[1] Some even attribute the Irish potato example to Robert Giffen himself.[2] This is odd since, if he made it, Giffen apparently failed to commit his suggestion to print.[3] In any case, Marshall's version of Giffen's paradox was presented in terms of *bread* (p. 132). With all this conjecture concerning an appearance of the rare upward-sloping demand curve in nineteenth-century Ireland, historical investigation of the question is long overdue.

We begin with an early statement of the "potato paradox" by Paul Samuelson:

> ...When the 1845 Irish famine greatly raised the price of potatoes, families who consumed a lot of potatoes merely because they were too poor to consume much meat might have ended up consuming *more* rather than less of the high-*P*[rice] potatoes. Why? Because now they had to spend so much on potatoes, the necessary of life, as to make it quite impossible to afford any meat at all and hence were forced to become even more dependent than before on potatoes. In brief, the substitution-effect was here overcome by the perverse income-effect applicable to a peculiar "inferior" good such as the potato, which tends to *decrease* in the poor man's budget when incomes rise. This *curiosum* is attributed to Sir Francis [*sic*] Giffen, a Victorian economist. [1964, p. 432, fn.]

More or less the same story is told in the more recent presentations; the blight reduced the supply of potatoes, increased the price, and consumers ate *more* potatoes. This cannot have occurred.

The most obvious problem with this hypothetical series of events is that, during the famine, the supply of potatoes in Ireland decreased, and there were not more but fewer potatoes available; the Irish people could not possibly have eaten more as a group. The fungus *phytophthora infestans* arrived in Ireland in 1845 after already damaging crops in America, where it first appeared, England, and on the Continent. In that year, it destroyed roughly one-half the crop in Ireland. The heavier rain in the summer of 1846 caused the potatoes to be infected with the fungus in the ground so that no more than one-tenth of the crop could be harvested (Cecil Woodham-Smith, 1962, pp. 91–93). Furthermore the spread of the blight seriously damaged crops in all neighboring countries, making it impossible to replace the lost potatoes through imports from other countries (Woodham-Smith, p. 118). After September 1846 until the harvest beginning in August of the following year, there were virtually no potatoes to be had at any price (Woodham-Smith, p. 109ff). The blight was not as serious in 1847 although, in part because seed potatoes had been eaten, the crop was still smaller than usual (Redcliffe Salaman, 1949, pp. 321–22). In 1848, the devastation was as widespread as in 1846.[4] Con-

---

*Department of Economics, Emory University, Atlanta, GA 30322.

[1] See, for example, Edgar Browning and Jacqueline Browning (1983, p. 82), Richard Leftwich (1982, p. 140), and Donald McCloskey (1982, p. 75).

[2] See, for example, Heinz Kohler (1982, p. 84) and Walter Nicholson (1978, p. 103).

[3] After a thorough search, George Stigler (1947) indicates that he found no evidence of such a suggestion by Giffen. A. R. Prest (1948) provides a quote which purportedly justifies the attribution to Giffen, but as Stigler (1948) indicates, it does not suggest a Giffen paradox.

[4] Unfortunately, the only year with reasonably thorough data in the above years is 1847. See B. R. Mitchell and Phyllis Deane (1962, p. 80). E. R. R. Green (1956) and Woodham-Smith indicate the quite substantial foundation for these statements.

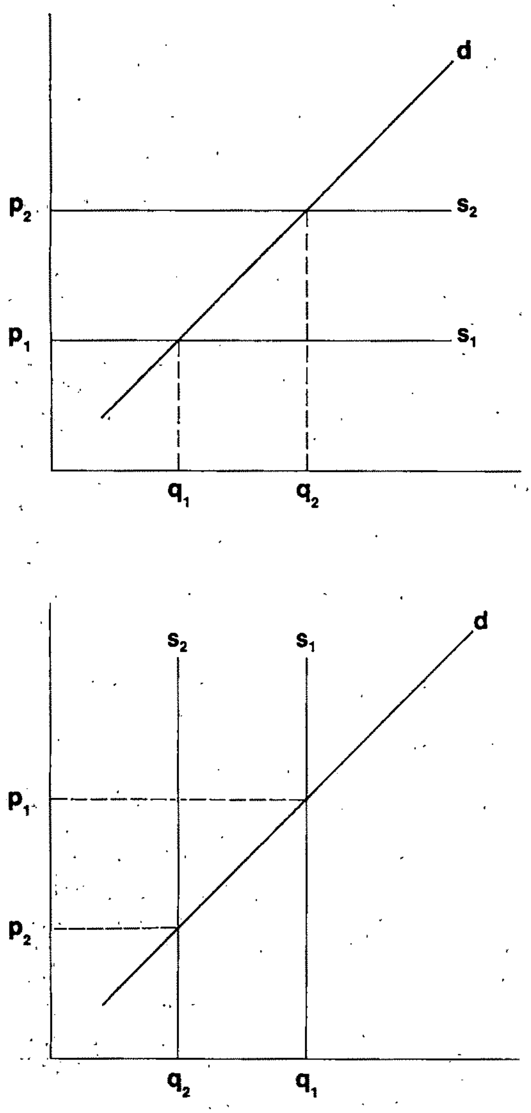


FIGURE 1. PRICE AND QUANTITY ADJUSTMENT FOR A GIFFEN GOOD

sumers in Ireland did not eat more potatoes during the famine.

Part of the difficulty of interpreting these events in the usual way is that conventional treatments focus on an individual demander's response to an exogenously changed price, while the Irish potato famine is a market experiment. Rural Ireland was virtually a closed economy in potatoes. The potato blight therefore shifted a very inelastic supply curve leftward. These different characterizations of the results are depicted in Figure 1. In the top panel price exogenously increases from P_1 to P_2, and demanders adjust quantity. In the bottom panel, which is more descriptive of the potato famine, supply shifts from S_1 and S_2, and price adjusts.

This raises a second problem with the conventional analysis. As pointed out above, that description has the Irish responding to a higher price by eating more potatoes. If the market demand for potatoes had been Giffen at this time, the Irish should have been responding to a lower market price by eating fewer potatoes! With an upward-sloping demand curve, a decrease in supply results in a lower, not a higher, price.[5] And for the market experiment posed by this case, it is the direction of the price change, not the quantity change, that illuminates the issue at hand.

This possibility is ludicrous: a large part of a country on the brink of starvation, and the market price of potatoes falls? The implication is difficult to refute, as there are no price quotations. There are two indirect pieces of evidence that are inconsistent with a lower price, however. Starving people logically would seek to maximize nourishment per unit of money spent on food. Potatoes are a relatively nourishing food; this is the reason that much of the population could subsist on them in nonfamine years and be relatively healthy. As a result, a fall in the price of potatoes relative to other foods implies a failure to optimize in this critical fashion. Only if the prices of other foodstuffs fell, would it be reasonable to expect the price of potatoes to have fallen. This implication of a fall in the price of potatoes therefore can be indirectly tested by changes in the prices of other foods. The prices of substitute foods increased substantially during the famine (Woodham-Smith, pp. 165–169; Mitchell and Deane, p. 488).

There is additional evidence that the price of potatoes did not fall during the famine. Relief agencies might be expected to optimize calories per unit of money with even greater disregard for other considerations than might starving peasants. The observed substitution out of potatoes during the famine period is therefore inconsistent with falling prices for these items. Prior to the famine, Irish workhouses provided a diet of large quantities of potatoes. During the famine, however, the food offered by English government and volunteer agencies was not potatoes, but imported American corn or corn meal (Thomas O'Neill, 1956; Woodham-Smith).

---

[5] For reasons indicated by L. A. Boland (1977) and William Dougan (1982), a decline in price is inconsistent with likely dynamics of the market.

The evidence presented above argues persuasively that the market demand for potatoes did not slope upward during the famine. Nonetheless, it might be argued that potatoes were a Giffen good for some consumers. In this case, the market demand might slope downward yielding a higher price at famine levels of supply to which some Giffen consumers responded by purchasing more. We cannot, of course, prove that no single consumer in all of Ireland behaved in this way. There is one group whose diet largely consisted of potatoes and whose expenditures on food, including home-produced food, were a large fraction of their income, namely the Irish peasantry. The standard rationale outlined in the Samuelson quote does not apply to this group.

For a good to be Giffen, some normal good must be displaced by the inferior good as the price rise lowers real income. Because the Irish peasantry subsisted almost exclusively on potatoes in good years, it is unclear what normal good would have been crowded out of the budgets to make way for more potatoes during the famine (had there been potatoes available). The Irish were poor even by contemporary European standards. Two-thirds of the Irish population worked in agriculture; 45 percent of the farms over an acre were five acres or less in size, and two-thirds were fifteen acres or less (R. M. Austin Bourke, 1965; Captain Larcom, 1843, pp. 348–51). In 1841, one-third of the houses in Ireland were one-room mud cabins, and over two-thirds were mud cabins with four or fewer rooms (T. W. Freeman, 1957, p. 151). The diet for much of the population consisted almost exclusively of potatoes together with oatmeal and milk. Even the comparatively more prosperous farmers who grew wheat, barley, and oats ate little of these higher-priced grains except in the summer months when potatoes would not keep. Little meat was eaten by any but the most prosperous (Freeman, pp. 69–70, 205; L. M. Cullen, 1981, pp. 147–48, 175). Besides growing potatoes, many of the poorer farmers raised a pig but this was cash crop, that is, the "gentleman who paid the rent" (Freeman, p. 56).[6]

There is a more important reason to conclude that this standard rationale does not apply to the farming poor. It is likely that potatoes were not inferior goods for the Irish peasantry during the famine years. Inferiority is necessary for a good to be Giffen. These farmers were producers as well as consumers, and the blight which destroyed their crops reduced their real incomes to near zero. Over broad income ranges, potatoes were inferior in prefamine Ireland.[7] At incomes in the starvation range, however, the income elasticity of food must be substantially positive. If potatoes had remained the cheapest source of nutrition during the famine, then their income elasticity would not have been negative. Nor can there be any doubt that the "hungry" portion of the income consumption curve is appropriate to our present inquiry. During the famine years, the situation was desperate almost beyond description. It has been estimated that roughly 1.1 million of a population of 9 million starved to death or died of diseases brought on by starvation from 1845 to 1851 (the first census after the famine), while another million fled the country (Joel Mokyr, 1979).

These observations raise doubts about the entire line of inquiry which led to the selection of the Irish potato as a likely candidate for Giffen status. It was adopted because in peasant economies such as Ireland a single staple occupies a dominant place in the budgets of these consumers; hence price changes can be expected to have important income effects. One problem with this approach is that, typically in these cases, the consumers are also producers of the staple. Income effects operating through the supply side of such economies confound attempts to isolate demand-side phenomena. In this case, for example, the potato blight in Ireland lowered real income to the point that potatoes were not inferior.

---

[6] See Bourke and Freeman (pp. 54–58). A somewhat larger proportion was reported in the Irish Census of 1841 (G. R. Porter, 1850).

[7] Across broad social classes, potato consumption in prefamine Ireland was lower for higher social classes (Cullen, pp. 140–92).

More importantly, typical treatments fail to reckon with the fact that in most cases these are closed economies in food. Conventional presentation of the Giffen paradox fixes money income and varies a single price. These price changes are simply imposed, and individuals are observed to quantity adjust. In peasant economies, it is quantity which varies exogenously and price which adjusts. Had the staple been produced externally and world conditions raised the price, then the scenario suggested in these presentations *might* have occurred.[8] We therefore conclude that the place to look for a Giffen good is not a peasant economy engaged in subsistence farming, but in poor communities that import most of their food. Has anyone considered rice in Singapore?

---

[8] In laboratory experiments with subjects that are consistent with the Slutzky-Hicks conditions, Raymond Battalio, John Kagel, and Carl Kogut (1983) have generated an increase in quantity demanded when price increases.

REFERENCES

Battalio, Raymond C., Kagel, John H. and Kogut, Carl, "Giffen Goods: They Really Exist," unpublished paper, Texas A&M University, 1983.

Boland, L. A., "Giffen Goods, Market Prices and Testability," *Australian Economic Papers*, June 1977, *16*, 72–85.

Bourke, R. M. Austin, "The Agricultural Statistics of the 1841 Census of Ireland: A Critical Review," *Economic History Review*, August 1965, *18*, 376–91.

Browning, Edgar K. and Browning, Jacqueline M., *Microeconomic Theory and Applications*, Boston: Little, Brown and Company, 1983.

Cullen, L. M., *The Emergence of Modern Ireland 1600–1900*, New York: Holmes and Meier, 1981.

Dougan, William R., "Giffen Goods and the Law of Demand," *Journal of Political Economy*, August 1982, *90*, 809–15.

Freeman, T. W., *Pre-Famine Ireland*, Manchester: Manchester University Press, 1957.

Green, E. R. R., "Agriculture," in R. Dudley Edwards and T. Desmond Williams, eds., *The Great Famine*, Dublin: Brown and Nolan Ltd. for the Irish Committee of Historical Sciences, 1956.

Kohler, Heinz, *Intermediate Microeconomics*, Glenview: Scott, Foresman & Company, 1982.

Larcom, Captain, "Observations on the Census of the Population of Ireland in 1841," *Journal of the Royal Statistical Society*, December 1843, *6*, 323–51.

Leftwich, Richard, *The Price System and Resource Allocation*. Hinsdale: Dryden Press, 1982.

McCloskey, Donald N., *The Applied Theory of Price*, New York: Macmillan, 1982.

Marshall, Alfred, *Principles of Economics*, 8th ed., New York: Macmillan, 1949.

Mitchell, B. R. and Deane, Phyllis, *Abstract of British Historical Statistics*, Cambridge: Cambridge University Press, 1962.

Mokyr, Joel, "The Deadly Fungus: An Econometric Investigation into the Short-term Demographic Impact of the Irish Famine: 1846–1851," in Julian L. Simon and Julie DeVanzo, eds., *Research in Population Economics*, Greenwich: JAI Press, 1979.

Nicholson, Walter, *Microeconomic Theory*, Hinsdale: Dryden Press, 1978.

O'Neill, Thomas P., "The Organization and Administration of Relief, 1845–52," in R. Dudley Edwards and J. Desmond Williams, eds., *The Great Famine*, Dublin: Brown and Nolan Ltd., 1956.

Porter, G. R., "Agricultural Statistics of Ireland," *Journal of the Royal Statistical Society*, Series A, 1850, *13*, 25–29.

Prest, A. R., "Notes on the History of the Giffen Paradox: Comment," *Journal of Political Economy*, February 1948, *56*, 58–60.

Salaman, Redcliffe N., *The History and Social Influence of the Potato*, Cambridge: Cambridge University Press, 1949.

Samuelson, Paul A. *Economics*, 5th ed., New York: McGraw Hill, 1964.

Stigler, George J., "Notes on the History of the Giffen Paradox," *Journal of Political Economy*, April 1947, *55*, 152–56.

______, "A Reply," *Journal of Political Economy*, February 1948, *56*, 61–62.

Woodham-Smith, Cecil, *The Great Hunger*, London: Hammish Hamilton Ltd., 1962.

Uncontrolled Prices in a Controlled Market: The Case of Rent Controls

By GEORGE FALLIS AND LAWRENCE B. SMITH*

In recent years there has been a proliferation of rent control legislation which places controls on most rental units, but exempts newly constructed units, vacated units, and/or high-priced units. The relationship between the rental price of the uncontrolled stock under these control regimes and the equilibrium rental price in the absence of controls is important in determining the beneficiaries and losers under controls and the impact of controls on the construction of new rental housing. Very little attention, however, has been directed either to the determination of this relationship in the housing market or in other controlled markets. Moreover, there has been little empirical analysis of the impact of rent or price controls with exemptions.

Previous discussions of rent controls with exemptions, such as that by Lionel Needleman (1965), asserted the rent on exempted units would be higher under a control regime than rents in the absence of controls, but offered no formal model of price determination. This assertion was challenged by J. R. Gould and S. G. B. Henry (1967) who, in a general study of price controls, demonstrated that the price of a substitute to the controlled good may rise or fall. They showed that the impact of controls on the price of a substitute depends upon the elasticities of the demand and supply curves for the two goods and also, importantly, upon the mechanism allocating the controlled good to households because that mechanism influences the amount of excess demand which "spills over" to the substitute. However, the Gould-Henry analysis cannot be directly applied to the housing market because it assumes households can simultaneously consume both the controlled good and the uncontrolled substitute; but, in the housing market, households consume in either the controlled or the uncontrolled sector, not both. Moreover, the two mechanisms examined for allocating the controlled good—a government rationing scheme dividing the goods evenly among consumers and a first-come, first-served procedure—are not very applicable to the analysis of most rent controls.[1]

This paper examines rent controls with exemptions by modeling the relationship between the rental price on uncontrolled rental units and the equilibrium price in the absence of controls, and by empirically investigating this relationship. Households face a binary choice: either to consume rental housing in the controlled or the uncontrolled market. Unlike price controls in other markets, the mechanism for allocating the controlled stock does not have to be arbitrarily assumed. Usually the mechanism is that the controlled stock is allocated to those households occupying it when the controls are imposed. Given the allocation mechanism, the nature of the shifts in the demand for housing determines how excess demand spills into the uncontrolled market. Section I presents a pricing model which demonstrates that uncontrolled rents can be higher or lower than rents in the absence of controls, but are

---

*Associate Professor of Economics, York University, Downsview, ON M3J1P3, and Professor of Economics, University of Toronto, 150 St. George Street, Toronto, ON M5S1A1, respectively. This research was funded by Social Sciences and Humanities Research Council of Canada, Research Award 451-81-2999 to Smith and a grant from Canada Mortgage and Housing Corporation to both authors. We thank Alan Heskin, Frank Mittlebach, and Kenneth Rosen, and participants in a seminar at York University for their helpfulness and comments.

[1] Steven Cheung (1975) raises the question of whether rent controls raise rents in the uncontrolled sector above the level that would have existed in the absence of controls, but focuses on the effect of controls on the demolition of existing controlled stock to make way for new construction of uncontrolled stock.

likely to be higher.[2]

Empirical investigation of rent controls with exemptions has always faced two difficulties: the development of a forecast of what rents would have been in the absence of controls, and the comparison of rents in the controlled and uncontrolled sector when the dwelling units in the two sectors were different. A data set from the Los Angeles area provides an opportunity to deal with both these difficulties. This is done in Section II which presents an empirical investigation of the effects of controls in the Los Angeles area and demonstrates that rents on exempted units are higher under rent controls than they would have been in the absence of controls.

I. The Model

A model is developed to determine the relationship between uncontrolled rents in a rent controlled environment and the equilibrium rents in the absence of controls. The analysis is developed for two rent control regimes, the first with new unit exemptions and the second with vacated unit exemptions.

A. *New Unit Exemption*

Consider an initial equilibrium in the rental housing market. The total quantity demanded of rental housing services, D, is the horizontal sum of the demands of the n households who rent housing in the community.[3] The demand D_i of the ith household is a function of rent R, a vector of other prices P, and household income M_i, as in equation (1). The quantity supplied, S, of rental housing services is a function of rent and a vector E of other variables, as set out in equation (2). The equilibrium condition, defined in equation (3) and shown graphically in Figure 1, generates an equilibrium rent R_c. The quantity of housing services exchanged is assumed to be proportional to the quantity of housing stock Q_c.[4]

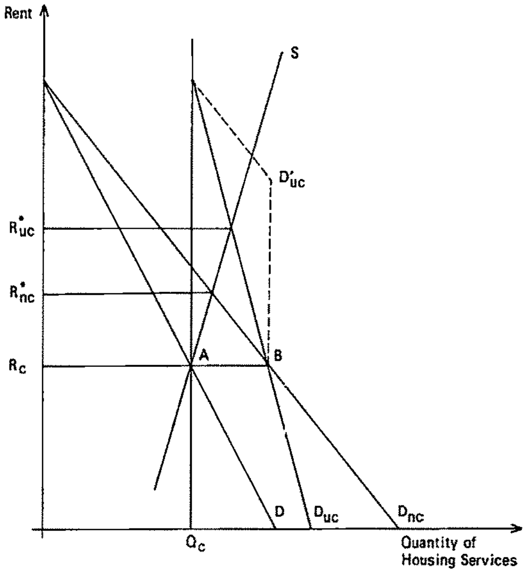


FIGURE 1. NEW UNIT EXEMPTION

$$(1)\qquad D=\sum_{i=1}^{n} D_i(R,P,M_i),$$

$$(2)\qquad S=f(R,E),$$

$$(3)\qquad \sum_{i=1}^{n} D_i(R,P,M_i)=f(R,E).$$

Assume now that an exogenous change, to

---

[2] The analysis in this paper is suitable with some modification for analyzing pricing in "black markets" when these markets emerge from price regulation. (See, for example, Edgar Browning and William Culbertson, 1974.) It may also shed light on other partially controlled markets such as natural gas and oil markets.

[3] The specification assumes that households do not shift between the ownership and rental markets. If such shifts occurred, the demand for rental housing would not be the horizontal sum over the n households who rent housing because the number of households in the rental market would change with the rent level. However, this assumption could easily be relaxed.

[4] The assumption that the quantity of housing services is proportional to the housing stock and the implicit assumption that units of stock and service are homogeneous are standard in housing analysis, and are discussed in Edgar Olsen (1969) and Richard Muth (1969). The notion of quality variation across dwellings or of a decline in quality of a dwelling unit through depreciation is treated as a change in the number of units of stock and service in a dwelling. The inclusion of a production function for housing services which uses stock and other inputs would not alter the results. All analysis would be done in terms of services, assuming controls actually regulate the service price. See also fnn. 5 and 7.

be explored below, shifts the market demand curve out and to the right, D_{nc} in Figure 1. Given that $\partial f/\partial R > 0$, rents would rise above R_c. Assume also that the government imposes rent controls on the existing stock Q_c, holding rents at R_c,[5] but exempts new housing to be constructed in the future.[6] There are now two markets: the controlled and the uncontrolled.

The demand for uncontrolled units may be represented as $D_{uc}(R_{uc}, R_c, P, M)$ where R_{uc} is the uncontrolled rent and M is the vector of incomes for all households. This function is not the horizontal sum of the demand functions of the group of households which secure housing in the uncontrolled market at the equilibrium because (as will be explored below) if R_{uc} changes, different households will enter the uncontrolled market. The condition determining equilibrium rent R_{uc}, in the uncontrolled sector is (4). New units will be built if R_{uc} is greater than R_c, and the quantity supplied is determined from the original supply function. For simplicity, it is assumed that the vector E is not influenced by the imposition of controls; this assumption is relaxed below.

$$(4)\quad D_{uc}(R_{uc}, R_c, P, M) = f(R_{uc}, E) - Q_c, \quad \text{for } R_{uc} > R_c.$$

The demand in the controlled market can be represented as $D_c(R_c, R_{uc}, P, M)$. The controlled supply, which for the moment is assumed to be perfectly inelastic, is Q_c.[7] Since the controlled market does not necessarily equilibrate, there may be excess demand at the controlled rent R_c.

Had controls not been imposed, the equilibrium rent R^*_{nc} would be established by (5) where the demand for rental housing in the absence of controls D_{nc} is defined as in (6). The total market demand in the world without controls is the sum of the demands in the controlled and uncontrolled markets; considering those demands as the horizontal sum of the demands of the j households in the controlled market and the $n-j$ households in the uncontrolled market determined for any given R_c and R^*_{uc}.

$$(5)\quad D_{nc}(R_{nc}, P, M) = f(R_{nc}, E),$$

$$(6)\quad D_{nc}(R_{nc}, P, M) = \sum_{i=1}^{j} D_i(R_{nc}, P, M) + \sum_{i=j+1}^{n} D_i(R_{nc}, P, M).$$

The pricing models are now complete except to specify the process by which households are allocated between controlled and uncontrolled markets. Demand in the uncontrolled market depends upon the mechanism used to ration the controlled stock. Under rent controls, the controlled stock is allocated to those occupying it when the controls are imposed. Given this rationing system, demand in the uncontrolled market depends upon the causes of the shift in the market demand curve (from D to D_{nc}). This shift could have a number of causes including an increase in the number of households through in-migration or an increase in the number of households through regrouping the existing population,[8] and an increase in demand due to changes in other prices or incomes.

Consider the case of in-migration to the community (or, alternatively, an increase in community headship rates) which causes the demand curve to shift. Since the existing households occupy the controlled stock, new entrants (new households) are confined to the uncontrolled market. The price determination process in this market is portrayed graphically in Figure 1.

---

[5] This assumes that government can actually control the price per unit of housing service. This is perhaps unrealistic because controls usually set the nominal gross rent but not the price per unit of service which is often affected by induced changes in output per dwelling unit. Mark Frankena (1975), and Smith and Peter Tomlinson (1981) discuss this issue.

[6] The imposition of rent controls could have been represented in the more traditional manner as a price below the initial equilibrium. The representation here, however, more realistically corresponds to circumstances which usually surround the imposition of controls and more clearly isolates the effects of a new unit exemption.

[7] This is not likely to be true because controls probably reduce maintenance and speed demolitions (Cheung, 1975). If housing services are produced using stock and other inputs, the assumption of inelastic supply becomes less tenable. The assumption is relaxed later in the text.

[8] Dwight Jaffee and Kenneth Rosen (1979) discuss the impacts of changing household size on the quantity of housing services demanded.

The vertical (rent) axis for the uncontrolled market is the vertical line through Q_c, and the horizontal axis which measures the quantity of newly constructed uncontrolled stock is the horizontal axis to the right of Q_c. The supply curve of new stock is AS, assuming controls do not affect E. The demand for uncontrolled housing is D_{uc} which intersects D_{nc} at B, because D_{nc} is the horizontal addition of D_{uc} and D. The uncontrolled rent is R^*_{uc} and is greater than rent in the absence of controls, R^*_{nc}. The amount of the difference depends upon the slopes of the demand and supply curves; the more inelastic the supply curve, and the more elastic the D_{nc} curve relative to the D curve, the greater is the difference. Both are higher than the controlled rent.[9]

It has been assumed that the variables in E were unchanged by controls, but this is not likely since the vector E contains expected future rents. Once controls have been imposed on the existing stock, there is a higher probability that controls will be imposed on new buildings, thus reducing expected future rents. This will cause the supply curve of new construction AS to shift up and probably also to become more inelastic. This further raises R^*_{uc} and increases the difference between R^*_{uc} and R^*_{nc}.

Now consider another cause of the increase in market demand: an increase in the income of existing residents, assuming there are no new entrants and no net household formation. In this case, the process allocating households between the controlled and uncontrolled markets is somewhat different. When demand increases and controls are imposed, all households occupy a controlled unit, but there is excess demand for housing services; AB in Figure 1. Each household faces a choice: either consume the existing quantity of housing services at the controlled price, or an unconstrained quantity of housing services at the higher uncontrolled price in the uncontrolled market. If any household shifts to the uncontrolled market, the choices facing the remaining households in the controlled market change slightly. Since a controlled unit has been vacated, there is an opportunity for some households to increase their housing consumption while remaining in the controlled market and enjoying the lower rent. The final equilibrium allocation of households between the two markets will thus depend on the sequence of departures from the controlled market and the sequence of moves in the controlled market.[10]

At one extreme, households with high incomes and strong preferences for housing (the aggregation of their demands forms the upper end of the D_{nc} curve) depart for the uncontrolled market, creating demand in the uncontrolled market represented by the dotted line D'_{uc}. Again, R^*_{uc} would be greater than R^*_{nc} and this would be the maximum possible difference for any supply curve. At the other extreme, few households depart the controlled market because most find utility is higher with a low rent and a housing constraint than with higher rent and no constraint. In this case, little demand spills into the uncontrolled market, excess demand in the controlled market is at a maximum, and, although R^*_{uc} and R^*_{nc} would be greater than R_c, it is possible that $R^*_{uc} < R^*_{nc}$.

It is likely that controls with a new unit exemption will cause $R^*_{uc} > R^*_{nc}$, since controls are likely to cause the supply curve to shift up and become more price inelastic, and since substantial demand is likely to spill into the uncontrolled market because of new entrants, net household formation, and the desire of some households to increase their housing consumption.

This analysis is of the relatively short run. In the very long run, all stock will be decontrolled because controlled stock will depreciate and disappear. Controls speed this disappearance. Landlords will reduce maintenance expenditure because such funds can earn a higher rate of return elsewhere. Landlords also have greater incentives to convert controlled stock to condominiums, or to nonresidential uses. As controlled stock declines (i.e., supply in the controlled market is less than perfectly inelastic over time), tenants in the controlled market consume less housing at the controlled rent and more will leave for the uncontrolled market.

---

[9] The model assumes that tenants have security of tenure and assumes landlords, tenants or prospective tenants do not make side payments of the sort discussed in Cheung (1974; 1975).

[10] The movement of a household to the uncontrolled sector could also be initiated because the household wishes to change location or the quantity of housing services consumed due to changes in family size.

B. *Vacated Unit Exemption*

Now consider the case of rent controls with vacancy decontrol provisions. Assume there is an initial uncontrolled equilibrium rent R associated with the market demand curve D_{nc} and supply curve s, shown in Figure 2. Assume further that controls are imposed at this rent, such that $R_c = R$, and that the supply curve of housing is perfectly inelastic.[11] Consider now the simple case of a vacating household who leaves the community and is replaced by a new household who enters the community and rents the decontrolled unit. The exit of the household shifts the market demand for controlled stock from D_{nc} to D_c and reduces the supply of controlled stock to Q_c. Since the household consumed AB before controls, this quantity is shifted from the controlled market and is now supplied in the uncontrolled market. Again, the diagram of the uncontrolled market can be superimposed on the original diagram. The vertical (rent) axis for the uncontrolled market is a vertical line through Q_c, and quantity is measured along the original horizontal axis to the right of Q_c. If the entering household has a demand function identical to the exiting household, the demand in the uncontrolled market is D_{uc} (the horizontal addition of D_{uc} and D_c is D_{nc}). The equilibrium rent in the uncontrolled sector would then be the same as the rent in the absence of controls and equal to R_c. If the entering household has a greater demand, the demand in the absence of controls becomes D'_{nc}, the demand in the uncontrolled market becomes D'_{uc} (again the sum of D'_{uc} and D_c is D'_{nc}), and the rent in the uncontrolled market becomes R^{**}_{uc}. The rent R^{**}_{uc} is greater than the rent in the absence of controls R^{**}_{nc}, and both are above the controlled rent R_c. If the replacement household has less demand, then controls are nonbinding. If the supply curve for uncontrolled housing is less than perfectly inelastic[12] (graphically, a positively sloped line through B), the price differential is less, but R^{**}_{uc} remains higher than R^{**}_{nc}.

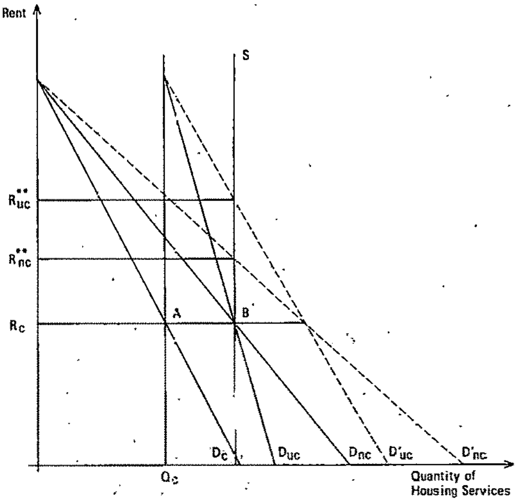


FIGURE 2. VACATED UNIT EXEMPTION

The above analysis assumes that the other existing households have no change in circumstances and do not compete for the vacant unit. However, if circumstances, such as income and household size, change for some existing households and they seek alternative housing in the uncontrolled market,[13] the analysis is similar to the case of replacement by greater demand, except there is a chain effect since the units vacated by the shifting households also become decontrolled.[14] Thus again, rent controls with vacancy decontrol will likely cause rents in the uncontrolled sector to be higher than rents in the absence of controls.

---

[11] Again, controls could have been represented as a price below the initial equilibrium. This representation and the assumption of inelastic supply help to isolate the phenomenon of vacancy decontrol from that of a new unit exemption.

[12] This can happen either by an increase in the supply of services from an existing dwelling unit or from new construction which is also exempt from controls.

[13] Eric Hanushek and John Quigley (1978) discuss factors affecting tenant mobility.

[14] We do not consider the case of departures to reduce the quantity of services consumed, since the

II. The Empirical Evidence

The preceding analysis suggests that the imposition of binding controls in one sector of the market causes rents in the uncontrolled sector to increase beyond the level they would have been without controls. This section empirically investigates this proposition using data from Los Angeles. A model of rent determination in the absence of controls is used to generate a forecast of rent increases for Los Angeles after two years of rent control. Actual rent increases in the controlled and uncontrolled markets in Los Angeles are then compared to the forecast, and uncontrolled rents are shown to have risen more than the forecast rent increase.

Los Angeles provides a unique set of data to test our hypothesis, as it recently enacted a second-generation form of rent control under which the annual allowable percentage increase in rent was restricted to 7 percent after May 31, 1978, subject to deregulation upon voluntary tenant movement,[15] and an exemption for newly constructed units.[16] Since a survey of controlled and uncontrolled rents was conducted two years after controls were introduced, Los Angeles provides a situation in which controlled and uncontrolled rents can be compared with forecasted rent at a time sufficiently close to the implementation of controls that we can be quite confident of the forecast. Moreover, since the vast majority of units that were decontrolled during the period were decontrolled through tenant movement, decontrolled units were often in the same buildings as controlled units. Consequently, the physical attributes of the controlled and uncontrolled units are likely to be very similar, and the data are likely to be free of the quality variations between the two markets that usually confound empirical analysis.

The model used to forecast rent is a variant of that developed by Smith (1974), and by Kenneth Rosen and Smith (1983). In this model, the demand for and the supply of rental housing determine the market vacancy rate V, and the deviation of this rate from the natural vacancy rate V^n determines the percentage change in the real rent $\dot{R}$, defined as the change in rent adjusted for the change in operating expenses $\dot{E}$, as

$$(7) \qquad \dot{R} = g(\dot{E}, (1/V - 1/V^n)).$$

This model was modified slightly to incorporate a lag structure in the rental price-adjustment process, and estimated in linearized form assuming the natural vacancy rate was constant over the estimation period. The estimated equation for Los Angeles using annual data from 1969 to 1978 is presented in equation (8), and confirms the appropriateness of the model.[17]

$$(8) \quad \underset{}{\dot{R}_t} = \underset{(3.30)}{-6.25} + \underset{(1.64)}{.078\dot{E}_t} + \underset{(4.12)}{34.09(1/V_t)} + \underset{(3.10)}{26.49(1/V_{t-1})}$$

$$\bar{R}^2 = .85; \; D\text{-}W = 1.89; \; F = 16.16; \; SER = .97$$

The estimated model was then used to forecast rent increases for Los Angeles in the absence of controls. The forecast indicates that rents would have increased 23.9 percent after two years in the absence of controls, on the assumption that the observed vacancy rate would have been unaffected by rent control. This assumption is not unreason-

---

higher price per unit of service in the uncontrolled market makes this unlikely, and since studies by Daniel Weinberg et al. (1981) and J. Doling (1976) have suggested that downward adjustments in demand are relatively unresponsive to declines in family size and traditional economic variables.

[15] The units become regulated again at the new rent level when a unit is reoccupied.

[16] For a discussion of rent control in Los Angeles, see W. A. V. Clark and Alan Heskin (1982), Clark et al. (1980), and Peter Rydell et al. (1981).

[17] The absolute value of the t-statistics are shown in parentheses. A second specification in which $\dot{R} = g(\dot{E}, V^n - V)$ was also estimated. This specification yielded slightly inferior results. However, forecasts based upon it were very similar to those based on equation (7).

TABLE 1—COMPARISON OF ACTUAL AND FORECAST RENT INCREASES AFTER TWO YEARS

	Actual Percentage Rent Increases	
	Rent Controlled Units	Decontrolled Units
District 1	13.6	38.5
2	14.8	59.4
3	13.3	45.2
4	13.3	44.1
5	14.1	47.5
6	12.6	43.1
Weighted Mean Increase[a]	13.7 ($n = 1435$)	46.2 ($n = 1137$)
Forecast Rent Increase	23.9	

Source: Actual rent increases were calculated from data in Clark and Heskin (p. 113).

[a] The number of observations in the sample is n.

able, since we are simulating forward only two years and it is unlikely that controls would have significantly affected the actual stock of rental housing during this short period.[18] If this assumption were incorrect, the observed vacancy rate would be lower than the rate that would have existed in the absence of controls and the assumption would cause an upward bias in the forecast rent. The assumption thus increases the robustness of our test.

The two-year forecast rent increase is compared in Table 1 with the actual rent increases in controlled and decontrolled units in six Los Angeles districts and the weighted mean increase for the entire area. The data were obtained from a study of rental housing in Los Angeles based on a survey of over 4,000 tenants in six districts with quite different demographic and socioeconomic profiles.[19] The base rent period for the survey is May 1978. Because the estimated equation and the forecast use annual observations, the forecast result may be considered to be centered as of the end of June, and thus corresponds reasonably well with the survey time period.

The results in Table 1 indicate that after two years, controlled rents had risen an average of 13.7 percent and uncontrolled rents had risen an average of 46.2 percent. These compare with the forecast increase of 23.9 percent. Consequently, the data confirm that rent controls effectively constrained rents on controlled units, but enabled larger rent increases on decontrolled units than would have occurred in the absence of rent controls. Furthermore, this conclusion is supported by the consistency of our findings since our result held for each district surveyed.

III. Conclusions

This paper extends the Gould and Henry analysis of the effect of controls on the price of a substitute good to consider the binary choice case relevant for the market in housing services, where a consumer purchases either in the controlled market or in the uncontrolled substitute market, but does not simultaneously purchase in both markets. The analysis considers a rationing process under which the controlled stock is allocated to households occupying it when controls are imposed and examines how shifts in the demand for housing affect the price of the uncontrolled substitute. The analysis indicates that under most conditions, rent controls with exemptions cause the rental price on uncontrolled units to exceed the equilibrium price in the absence of controls.

The conclusion was supported by an empirical investigation of pricing under rent controls in Los Angeles. Actual rent increases on controlled and uncontrolled units were compared with a forecast of rent increases in the absence of controls generated by a model of rent determination. The comparison indicated that after two years of rent control, rents on uncontrolled rental units were significantly above the forecast rent while rents on controlled units were below the forecast rent.

REFERENCES

Browning, Edgar K. and Culbertson, William Patton, "A Theory of Black Markets Under

---

[18] The vacancy rate is likely to be somewhat lower because rents do not rise to ration demand in the controlled sector.

[19] For a discussion of the survey and the data, see Clark and Heskin, and Clark et al.

Price Control: Competition and Monopoly," *Economic Inquiry*, June 1974, *12*, 175–89.

Cheung, Steven N. S., "A Theory of Price Control," *Journal of Law and Economics*, April 1974, *17*, 53–71.

______, "Roofs or Stars: The Stated Intents and Actual Effects of a Rents Ordinance," *Economic Inquiry*, March 1975, *13*, 1–21.

Clark, W. A. V. and Heskin, Allan D., "The Impact of Rent Control on Tenure Discounts and Residential Mobility," *Land Economics*, February 1982, *58*, 109–17.

______, ______, **and Manuel, Louise**, *Rental Housing In the City of Los Angeles: A Report of a Market Analysis and Statistical Research on Dwellings in the City of Los Angeles*, Institute for Social Science Research, University of California-Los Angeles, May 1980.

Doling, J., "The Family Life Cycle and Housing Choice," *Urban Studies*, February 1976, *13*, 55–58.

Frankena, Mark, "Alternative Models of Rent Control," *Urban Studies*, October 1975, *12*, 303–08.

Gould, J. R. and Henry, S. G. B., "The Effects of Price Control on a Related Market," *Economica*, February 1967, *34*, 42–49.

Hanushek, Eric A. and Quigley, John M., "An Explicit Model of Intra-Metropolitan Mobility," *Land Economics*, November 1978, *54*, 411–29.

Jaffee, Dwight M. and Rosen, Kenneth T., "Mortgage Credit Availability and Residential Construction," *Brookings Papers on Economic Activity* 2:1979, 333–76.

Muth, Richard, *Cities and Housing*, Chicago: University of Chicago Press, 1969.

Needleman, Lionel, *The Economics of Housing*, London: Staples Press, 1965.

Olsen, Edgar O., "A Competitive Theory of the Housing Market," *American Economic Review*, September 1969, *59*, 612–22.

Rosen, Kenneth T. and Smith, Lawrence B., "The Price-Adjustment Process for Rental Housing and the Natural Vacancy Rate," *American Economic Review*, September 1983, *73*, 779–86.

Rydell, Peter C. et al., *The Impact of Rent Control on the Los Angeles Housing Market*, Santa Monica: Rand Corporation, 1981.

Smith, Lawrence B., "A Note on the Price-Adjustment Mechanism for Rental Housing," *American Economic Review*, June 1974, *64*, 478–81.

______ **and Tomlinson, Peter**, "Rent Controls in Ontario: Roofs or Ceilings?," *American Real Estate and Urban Economics Association Journal*, Summer 1981, *9*, 93–114.

Weinberg, Daniel H., Friedman, Joseph and Mayo, Stephen K., "Intra-Urban Residential Mobility: The Role of Transactions Costs, Market Imperfections and Household Disequilibrium," *Journal of Urban Economics*, May 1981, *9*, 332–48.

The Emergence of Countercyclical U.S. Fertility: Note

By WALTER KRÄMER AND KLAUS NEUSSER*

In a recent paper in this *Review* (1979), William Butz and Michael Ward (hereafter B-W) suggest the following model for postwar U.S. fertility rates:

$$(1)\qquad \ln B = \gamma_0 + \gamma_1 K \cdot \ln Y_m + \gamma_2 \ln Y_m + \gamma_3 K \cdot \ln W_f ,$$

where B is the probability that a couple will have a child in a given year (replaced for estimation with its population analogue, the fertility rate), K is the fraction of families with employed wives, Y_m is the husband's income, W_f is the wife's wage, and ln denotes natural logarithms. Hypothesizing a positive association between male income and fertility, Butz and Ward suggest that both γ_2 and $\gamma_1 + \gamma_2$ should be positive, which is supported by their regression results.

The purpose of this note is to point out that regression estimates of (1) depend to a considerable extent on the monetary unit chosen to express Y_m and W_f, and that a change of this unit leads to different estimates for the parameters. In particular, negative estimates for both γ_2 and $\gamma_1 + \gamma_2$ can be obtained from the B-W data set through suitable choice of the monetary unit.

From the peculiar way in which the right-hand variables enter the equation, it is easily seen that different monetary units lead to different regression spaces, both for ordinary least squares (*OLS*) and for two-stage least squares (*2SLS*) estimation. Following Butz and Ward, let Y_m and W_f denote the income and wage variables, respectively, as measured in constant 1964 dollars. Allowing for alternative monetary units, (1) may more generally be written as

$$(2)\qquad \ln B = \gamma_0 + \gamma_1 K \cdot \ln(\lambda Y_m) + \gamma_2 \ln(\lambda Y_m) + \gamma_3 K \cdot \ln(\lambda W_f).$$

Table 1 reports *2SLS* estimates of (2) for various values of λ and for the female age group, 20–24. Similar results hold for other age groups as well. As in Butz and Ward, $K \cdot \ln(\lambda Y_m)$ and $K \cdot \ln(\lambda W_f)$ are viewed as endogenous, with $\ln Y_m$, $\ln Y_{m-1}$, $\ln W_f$, and $\ln W_{f-1}$ being the instruments. For $\lambda = 1$, results correspond to the first column in Table 1 of Butz and Ward.[1] However, multiplying the monetary unit by 10 ($\lambda = 0.1$) yields a negative estimate for $\gamma_1 + \gamma_2$, and a negative estimate for γ_2 results when λ approaches either 0 or ∞.

To obtain the limiting *2SLS* regression estimates as $|\ln \lambda| \rightarrow \infty$, let an asterisk denote right-hand variables as fitted to the instruments (which are easily seen not to depend on λ), let a tilde denote *2SLS* regression estimates, and consider

$$(3)\qquad \widetilde{\ln B}^{(\lambda)} = \tilde{\gamma}_0^{(\lambda)} + \tilde{\gamma}_1^{(\lambda)} [K \cdot \ln(\lambda Y_m)]^* + \tilde{\gamma}_2^{(\lambda)} \ln(\lambda Y_m) + \tilde{\gamma}_3^{(\lambda)} [K \cdot \ln(\lambda W_f)]^* ,$$

where the superscript λ underlines the dependency of the regression results on the scale parameter. Using

$$[K \cdot \ln(\lambda Y_m)]^* = \ln \lambda \cdot K^* + [K \cdot \ln Y_m]^* ,$$

$$[K \cdot \ln(\lambda W_f)]^* = \ln \lambda \cdot K^* + [K \cdot \ln W_f]^* ,$$

---

*Institute for Advanced Studies, Stumpergasse 56, A-1060 Vienna. We are indebted to Robert Coen, Warren Sanderson, and an anonymous referee for helpful suggestions and comments. The IAS-SYSTEM was used for computation and estimation.

[1] Estimates are computed from the data as reported in Butz and Ward (1977), since the data for the 1979 article have not been made available to us. The time-series in the 1977 paper are identical to those in the 1979 article, except that the 1975 data are missing. This leads to slightly different parameter estimates, without affecting the thrust of the argument.

TABLE 1—*2SLS* ESTIMATES OF THE B-W FERTILITY EQUATION FOR VARIOUS VALUES OF λ[a]

Independent Variable	λ										
	0	e^{-100}	e^{-10}	0.01	0.1	1	10	100	e^{10}	e^{100}	∞
$K\cdot\ln(\lambda W_f)$	−1.62	−1.76	−1.77	−1.53	−1.56	−1.58	−1.59	−1.59	−1.60	−1.89	−1.62
$K\cdot\ln(\lambda Y_m)$	1.62	1.96	14.04	−2.83	−1.21	−0.46	−0.02	0.26	0.66	1.72	1.62
$\ln(\lambda Y_m)$	−0.03	−0.18	−5.28	1.84	1.16	0.85	0.66	0.54	0.38	−0.83	−0.03
Intercept	$-\infty$	−8.84	−0.63	0.33	0.28	0.30	0.34	0.39	0.52	16.30	∞
R^2	0.919	0.915	0.858	0.928	0.926	0.924	0.923	0.923	0.922	0.912	0.919

[a] Female age group, 20–24. The data are from Butz and Ward (1977, p. 44). The results in column $\lambda = 1$ correspond to column 1 in Table 1 from Butz and Ward (1977, p. 8).

this can (after some algebra) be written as

$$\begin{aligned}(4)\quad \ln B^{(\lambda)} = {} & \left(\tilde{\gamma}_0^{(\lambda)} + \tilde{\gamma}_2^{(\lambda)} \ln\lambda\right) \\ & + \tilde{\gamma}_1^{(\lambda)}\left\{[K\cdot\ln Y_m]^* - [K\cdot\ln W_f]^*\right\} \\ & + \tilde{\gamma}_2^{(\lambda)}\ln Y_m + \left[\left(\tilde{\gamma}_1^{(\lambda)} + \tilde{\gamma}_3^{(\lambda)}\right)\cdot\ln\lambda\right] \\ & \times\left\{K^* + [K\cdot\ln W_f]^*/\ln\lambda\right\}.\end{aligned}$$

The last regressor in (4) tends to K^* as $|\ln\lambda| \to \infty$ and the limiting regression estimates in (4) can thus be obtained by regressing $\ln B$ on $\{[K\cdot\ln Y_m]^* - [K\cdot\ln W_f]^*\}$, $\ln Y_m$ and K^*. The respective coefficients as computed from the B-W data are 7.72 (the constant), 1.62, −0.03, and −17.55, which in view of (4) implies $\tilde{\gamma}_1^{(\lambda)} \to 1.62$, $\tilde{\gamma}_1^{(\lambda)} + \tilde{\gamma}_3^{(\lambda)} \to 0$, $\tilde{\gamma}_0^{(\lambda)} \to \pm\infty$, and $\tilde{\gamma}_2^{(\lambda)} \to -0.03$ as $|\ln\lambda| \to \infty$.

Table 1 shows that the nonhomogeneity of the B-W model leads to regression results that contradict the principal message of the original paper and thus cannot be viewed as a minor byproduct of some simplifications along the way to (1). By neglecting "small" terms in their original derivation (B-W, 1979, p. 320), they implicitly impose the constraint $dB/dK = 0$ and rule out K as a separate regressor. With K included in (1), the regression space for both *OLS* and *2SLS* as well as the respective parameter estimates for γ_i ($i = 1,\dots,3$; not the constant and the K coefficient) no longer vary with λ. The unit problem thus almost disappears, but the resulting estimates are not in line with Butz and Ward's underlying theory, either. The *2SLS* estimates yield (asymptotic t-values in parentheses)

$$\begin{aligned}(5)\quad \ln B = {} & \underset{(.45)}{35.28} + \underset{(.42)}{9.31}K\cdot\ln Y_m - \underset{(.35)}{3.28}\ln Y_m \\ & - \underset{(1.82)}{1.71}K\cdot\ln W_f - \underset{(.44)}{82.68}K,\end{aligned}$$

where γ_2 is negative, contrary to what Butz and Ward suggest.[2] Assuming that their theory is correct, therefore, it certainly cannot be cast in a form such as equation (1).

---

[2] This result treats K, $K\ln Y_m$ and $K\ln W_f$ as endogenous variables and uses $\ln Y_m$, $\ln Y_{m-1}$, $\ln W_f$ and $\ln W_{f-1}$ as instruments. Additional regressions were run with K_{-1} as another instrument, with equally implausible results. The *OLS* estimates also contradict the B-W hypothesis. We would like to point out, however, that the regression results are numerically unstable when K is included, due to near collinearity of the regressors. This is also reflected in very low t-values.

REFERENCES

Butz, William P. and Ward, Michael M., "The Emergence of Counter-Cyclical U.S. Fertility," Report No. R-1605-NIH, Santa Monica: Rand Corporation, 1977.

______ **and** ______, "The Emergence of Countercyclical U.S. Fertility," *American Economic Review*, June 1979, *69*, 318–28.

The *q* Theory of Investment with Many Capital Goods

By DAVID E. WILDASIN*

The "*q*" theory of investment, which relates investment to the ratio of market to replacement value of capital, has attracted considerable attention in a recent series of papers. For instance, a *q* variable has been used as an independent variable in empirical investment equations estimated by George von Furstenberg (1977), von Furstenberg et al. (1980), Burton Malkiel et al. (1979), and others. These papers, however, leave somewhat unclear the theoretical rationale for using *q* as an investment determinant. This has motivated work by Hiroshi Yoshikawa (1980), Lawrence Summers (1981), Michael Salinger and Summers (1981), and Fumio Hayashi (1982), who show that under certain conditions the rate of investment of a share-value-maximizing firm is indeed a function of *q*. This is an important result because it shows that investment equations with *q* an independent variable are not *ad hoc* constructions. Rather, they are grounded in a theory of the firm with an appealing behavioral hypothesis, viz, value maximization.

An important assumption underlying this research is that capital can be treated as a homogeneous good. Of course, this is an extremely common assumption in the analysis of investment, and is not particularly more bothersome in the *q* theory context than elsewhere. Nonetheless, there are certain situations where it may be desirable or even essential to be able to study investment disaggregated by type of capital good. Thus, the purpose of this paper is to examine whether and how the *q* theory can be extended to this more general case.[1]

Intuitively, one would expect some difficulty with the *q* theory in the many-capital-good context, as already noted by James Tobin and William Brainard (1977, p. 243) and by Salinger and Summers (p. 12). One way to see why is to recall the Tobin-Brainard distinction between "marginal" and "average" *q*. As Hayashi writes,

> The "*q*" theory...is not operational as long as *q* is not observable. Remember that *q*, which we call marginal *q*, is the ratio of the market value of an *additional* unit of capital to its replacement cost. What we can observe is *average q*, namely the ratio of the market value of *existing* capital to its replacement cost. [p. 214]
>
> If we knew marginal *q*, then econometric implementation of the "*q*" theory would be quite straightforward. Unfortunately, however, marginal *q* is not directly observable. What we can (in principle) observe is average *q*....

---

*Associate Professor, Department of Economics, Indiana University, Bloomington, IN 47405. This note arose from discussions with participants in the Spring 1982 Public Finance Seminar at Indiana University, for which I am grateful. Revisions were undertaken during my year at Queen's University. A referee also provided useful comments.

[1] Two examples will illustrate the significance of the extension to the many-capital goods case (which Salinger and Summers identify as one of the "two most important area[s] for further investigation," p. 47). First, relative price shocks or other changes in the economy may reduce the value of existing capital, may make new investments in some sectors of the economy unattractive, and may make some new investments very attractive, causing previously aligned marginal and average *q*s to diverge. The 1973 change in energy prices, cited by Salinger and Summers (p. 12), is the classic example. Numerous studies (see, for example, Martin Baily, 1981) suggest that this shock has effectively rendered a substantial amount of existing capital obsolete, simulanteously creating important incentives for new investment.

Second, consider the new "15-5-3" tax rules for depreciation of different categories of capital goods (cars and light trucks, other vehicles and equipment, structures, etc.). Whether one is interested in the effects of this policy change on investment disaggregated by these categories or just the effect on aggregate investment, there is little reason to believe that this policy change is adequately described by a change in average *q*, or by any other single variable. Note that the inadequacy of a single capital good model in this case is inherent in the complexity of the policy change one wishes to analyze. Short of ignoring the policy question itself, there is no way to finesse this difficulty by assuming a single capital good.

> [P]eople are busying themselves regressing investment on average q. Researchers should feel uneasy about doing this, unless they are sure that average q and marginal q are practically the same thing. The...[main theorem for the case of a competitive firm] states that marginal q and average q are essentially the same.... [p. 218]

In other words, marginal q is the fundamental determinant of investment because it shows how much increase in market value accompanies a dollar's worth of investment, while the actual stock market value of the firm reflects the profitability of existing total capital. Hayashi's main contribution is to establish the connection between the two. But with many capital goods, can a similar link be forged? A seemingly likely difficulty is that at any particular moment, a firm that employs a variety of capital goods will find that some kinds of investments may be highly attractive, others less so, and still others may be not worth undertaking at all. If so, there will be a variety of marginal qs, one for each capital good. Obviously, they cannot all be equal to the same "observable" average q. The link between average q and the investment-determining marginal qs may therefore be broken.

To explore this intuition formally, Section I presents a model similar to those of Hayashi, Summers, and Salinger-Summers, but which allows for an arbitrary number of capital goods. Section II briefly recapitulates the main q theory results for the special case of a single capital good. Section III turns to the main issues and, as expected, confirms the above intuition: in general, total investment in many capital goods cannot be expressed as a monotonic function of q. Necessary and sufficient conditions for this to be possible, involving mainly restrictions on adjustment cost functions, are presented, as well as particular examples of such functions.

Section III does not only contain this bad news, however. It also presents an important positive result. I derive a relationship between q and the *vector* of investment in the various capital goods, which shows how q is uniquely a function of this vector through the structure of the adjustment cost function(s). This means that regression equations with the q variable on the *left*-hand side and the vector of investment on the *right*-hand side can be used to estimate the underlying technology of investment as reflected in the parameters of the adjustment cost function(s). Since the relationship between the vector of investment and q is many-to-one rather than one-to-one, it cannot generally be inverted to express investment as a function of q. (This indeed is the bad news noted above.) Thus, we cannot use estimates from such an equation to predict total investment, conditional on q. Despite this limitation, however, the q theory approach still proves of value if one's interest is to determine structural parameters governing the investment process.

I. The Model

For the sake of simplicity in exposition, I assume the complete absence of personal or corporate taxes, and of debt finance. The results generalize to a model which includes those of Hayashi and Summers as special cases, however.[2]

The first step in constructing the model is to integrate the firm's share value equilibrium condition

$$(1)\qquad (\rho+\pi_g)V=\dot{V}+\mathrm{Div},$$

from $t=0$ to $t=\infty$, invoking a boundedness condition, to solve for the equilibrium value of shares outstanding at $t=0$:

$$(2)\qquad V(0)=\int_0^\infty \mu \mathrm{Div}\,dt,$$

where

$$(3)\qquad \mu(t)=\exp\left(-\int_0^t(\rho+\pi_g)\,ds\right).$$

Here and below all variables should be considered time dependent unless otherwise stated, although time arguments will be suppressed except where they are essential. We have V = aggregate nominal value of shares outstanding; ρ = exogenously given real rate of return on equity; π_g = exogenously given rate of change of general price level; Div = nominal dividends paid out, and a dot above a variable denotes a time derivative.

---

[2] A version of this paper containing this extension is available from the author on request.

At any moment of time, the firm receives a flow of cash from sales of output, and expends cash on hiring variable factors and on investment, including adjustment costs. Any remaining cash flow is paid out as dividends, which are assumed always to be positive (or at least the nonnegativity constraint is ignored). I assume that output is a linear homogeneous function $F(K, L)$ defined over the vector $K = (K_1, \dots, K_m)^T$ of stocks of m capital goods and the vector $L = (L_1, \dots, L_n)^T$ of flows of n variable factors. The price of output is p_0, $w = (w_1, \dots, w_n)$ is the vector of variable input prices, and $p = (p_1, \dots, p_m)$ is the vector of capital goods prices, all in nominal terms. The stock of capital good j evolves according to

$$(4) \qquad \dot{K}_j = I_j - \delta_j K_j,$$

where I_j is gross investment and δ_j is the (not necessarily time invariant) proportional rate of depreciation. The firm incurs adjustment costs when undertaking investment, which is assumed to take the form either of lost output or of waste of the good being invested. Specifically, let $\Gamma(I, K)$ be a linear homogeneous adjustment cost function showing units of lost *output* when the vector of investment is $I = (I_1, \dots, I_m)$ and the stock of capital is K, and let $\Phi_j(I_j, K_j)$ be a linear homogeneous adjustment cost function showing units of *good j* lost from investment I_j given capital K_j. Both kinds of adjustment costs appear in the literature; presumably one or the other would be more appropriate in particular cases. Actually, still more general forms are conceivable, but consideration of them would not change the results of this paper. Here I shall carry both kinds of adjustment costs through most of the formal analysis, focusing later on one or the other special case as convenient. In addition to linear homogeneity, the adjustment-cost functions are assumed to be twice continuously differentiable; the $m \times m$ matrix $\partial^2\Gamma/\partial I_j \partial I_{j'}$ is assumed positive definite; $\partial^2\Phi_j/\partial I_j^2$ is assumed positive; $\Phi_j(0, K_j) = \Gamma(0, \dots, 0; K) = 0$; and $\Phi_j(I_j, K_j) \geq 0 \leq \Gamma(I, K)$, for all I, K.

Given these assumptions, we can write

$$(5) \quad \text{Div} = \left(p_0[F(K, L) - \Gamma(I, K)] - wL\right) - \Sigma_j\left[p_j I_j + p_j \Phi_j(I_j, K_j)\right],$$

or, substituting into (2),

$$(6) \quad V(0) = \int_0^\infty \left\{p_0[F(K, L) - \Gamma(I, K)] - wL - \Sigma_j\left[p_j I_j + p_j \Phi_j(I_j, K_j)\right]\right\} \mu \, dt.$$

The firm's problem is to choose paths for L and I to maximize (6) subject to (4). Associating auxilary variables λ_j with (4), we have the following necessary conditions for an optimum, letting subscripts on F, Γ, and Φ_j denote partial derivatives:

$$(7) \qquad p_0 F_{L_k} - w_k = 0, \text{ all } k, t;$$

$$(8) \quad \lambda_j - \left[p_0 \Gamma_{I_j} + p_j\left(1 + \Phi_{I_j}\right)\right]\mu = 0, \text{ all } j, t;$$

$$(9) \quad -\dot{\lambda}_j = \left[p_0\left(F_{K_j} - \Gamma_{K_j}\right) - p_j \Phi_{K_j}\right]\mu - \lambda_j \delta_j, \quad \text{all } j, t;$$

$$(10) \qquad \lim_{t \to \infty} \lambda_j(t) K_j(t) = 0, \text{ all } j.$$

II. The Single Capital Good Case

Let us first review the principal results of the q theory for the special case $m = 1$. We begin by interpreting (8) as an equation determining the optimal I (dropping the capital good subscript) as a function of the shadow price of capital, λ, and other variables. The problem with this equation is that it cannot be estimated as it stands, since λ is unobservable. However, Hayashi proves that

$$(11) \qquad \lambda(0) K(0) = V(0),$$

by a method of proof explained in the next section. This equation says that the shadow value of the initial capital stock is equal to the current stock value of the firm. What is crucial about (11) is that it allows one to solve for $\lambda(0)$ and thus to rewrite (8) in a form more suitable for empirical application.

Specifically, by the homogeneity of Γ, note that $\Gamma(I,K)=K\Gamma(I/K,1)\equiv K\gamma(I/K)$, so that $\Gamma_1=\gamma'$ is the lost output per unit increase of investment. Similarly, one may define $\phi(I/K)\equiv K\Phi(I/K,1)$. Then, recalling that p is the price of the capital good, $(p_0\gamma'+p\phi')/p$ is the adjustment cost, in dollars, per dollar's worth of investment. Using (11) in (8), we have an equation for I/K:

$$(12)\quad [p_0\gamma'(I/K)+p\phi'(I/K)]/p = V/pK-1\equiv Q$$

or

$$(13)\qquad I/K=h(Q),$$

where $h(\cdot)=[(p_0\gamma'(\cdot)+p\phi'(\cdot))/p]^{-1}$ and where Q is Tobin's q minus 1.

The existence of the inverse defining $h(\cdot)$ in (13) is assured by the convexity of γ and ϕ. Thus, the rate of investment—or (multiplying both sides of (13) by K) the level of investment—is an increasing function of Q. Depending on the form of the adjustment costs (i.e., whether they involve lost output, with $\Phi\equiv 0$, or wasted investment goods, with $\Gamma\equiv 0$), we can use (13) in a regression of I/K on Q to identify parameters characterizing the adjustment cost technology, as in Summers and Hayashi.

Notice that the crucial contribution of the q theory in the single capital good case is that it permits one to replace the unobservable shadow value λ by V/K, from (11), which allows one to write the first-order condition for investment, (8), in a form that can be estimated.[3]

III. The General Case

A. *Unobservability of the Shadow Values of Capital*

Let us return to the general many capital good case. From (8) one sees that each I_j depends on the own-shadow price λ_j and, through the function Γ_{Ij}, on the amount of investment in other capital goods, as well as on various parameters. In fact, (8) can now be seen as a system in which the investment levels $\{I_j\}$ are all jointly determined by shadow values for all capital goods and by other parameters. Of course, this system cannot be used as it stands to estimate a system of investment equations because the λ_js are not observed. The natural question is whether the λ_js can be eliminated, as in the single good case, by using the market's forward-looking valuation of capital as reflected in the stock value.

We can proceed to answer this question by applying Hayashi's essential insight in the many-capital good case, and write, by (10),

$$(14)\quad \lambda_j(0)K_j(0)=\lambda_j(0)K_j(0)-\lim_{t\to\infty}\lambda_j(t)K_j(t) = -\int_0^{\infty}(\dot{\lambda}_jK_j+\lambda_j\dot{K}_j)\,dt.$$

Using (7), (8) and (9), we have

$$(15)\quad -\dot{\lambda}_jK_j-\lambda_j\dot{K}_j=\Big[p_0\big(F_{K_j}K_j-\Gamma_{K_j}K_j-\Gamma_{I_j}I_j\big)+\Sigma_k\big(p_0F_{L_K}-w_k\big)L_k - p_jI_j-p_j\big(\Phi_{I_j}I_j+\Phi_{K_j}K_j\big)\Big]\mu.$$

---

[3] Hayashi and Summers stress the "observability" of average q (see the quotation in the opening discussion attributed to Hayashi, for example), and indeed (11) does appear to confirm that an unobservable shadow value can be expressed in terms of currently observable variables. Unfortunately, this is not possible in a world with taxes. As discussed in the unpublished version of this paper, a "tax-adjusted" Q of the sort used by Hayashi and Summers cannot be estimated without knowing the expected time path of all personal and corporate tax policy parameters (with the exception of the investment tax credit), and expected future tax policy is not directly observed. To get around this problem, Hayashi, Summers, and Salinger-Summers assume that agents held static expectations on tax policy parameters in every sample period, and on this basis they are able to compute Q-type variables which are then used in their regressions; unfortunately their estimates are thus dependent on the maintained hypothesis of static expectations. It is regrettable that the true observability of Q that obtains in the simple taxless world of this paper does not carry over to a world with taxes, for if a tax-adjusted Q really were observable it would be possible (in a single capital good world) to estimate policy-invariant parameters of an investment equation.

At this point the derivation of the tax-adjusted Q breaks down. In the single capital-good case, homogeneity of F and Γ (application of Euler's theorem) allows one to use (6) and (14) to produce (11). By contrast, with many capital goods, integrating and summing over j (using (6) and (14)) yields

$$\Sigma_j\lambda_j(0)K_j(0)=V(0). \tag{16}$$

This equation, the many capital-good analogue of (11), has the same intuitive interpretation: the shadow value of the initial capital stock is equal to the current stock market value of the firm. However, there is also a fundamental difference between the two cases: whereas (11) allows one to estimate the single shadow value of capital from $V(0)$ and $K(0)$, (16) reveals that $V(0)$ is an initial capital-weighted sum of the multiple shadow prices, λ_j, and cannot be used to identify any *one* shadow price. In short, the many marginal qs, represented by the λ_js, cannot each be estimated by the one average q, represented by $V(0)$, confirming the intuitively obvious claim made in the introduction.

B. *Identification of Technological Parameters*

The above limitation certainly does not mean that a q-type variable cannot be useful for empirical work, however. Notice from (8) that (at $t=0$)

$$\left(p_0\Gamma_{I_j}+p_j\Phi_{I_j}\right)K_j=\lambda_jK_j-p_jK_j. \tag{17}$$

Summing over j using (16), and defining the weights $\omega_j=p_jK_j/pK$, one has

$$\Sigma_j\left(\left(p_0\Gamma_{I_j}+p_j\Phi_{I_j}\right)/p_j\right)\omega_j = V/pK-1=Q. \tag{18}$$

Note the similarities and differences between (12) and (18). The latter is clearly a natural generalization of the former, with a q-type variable on the right and a marginal adjustment cost term on the left. In the case of (12), the left-hand side is simply the marginal adjustment cost per dollar's worth of investment in the single capital good. In (18), the corresponding expression is a weighted sum of the marginal adjustment cost per dollar of investment for each of the m capital goods. The weights are simply the value share of each capital good in the firm's initial capital stock.

If one had observations on individual I_js and K_js, (18) could be used as the basis of an econometric model in which Q is regressed on the left-hand side variables, which are functions of the vector (I,K). Given a time-series for all relevant variables, one could estimate the parameters of the adjustment cost function(s), in either of the two specifications we have permitted (i.e., "lost output" or "lost capital").

As an example, one might suppose that adjustment costs take the form of wasted capital, and that each Φ_j is a quadratic:

$$\Phi_j(I_j,K_j)=(\alpha_j/2)(I_j/K_j)^2,\ \text{all}\ j.$$

Then, by (18), one has

$$Q=\Sigma_j\alpha_j\omega_j\left(I_j/K_j^2\right) \tag{19}$$

which, if formulated as a linear regression model, would yield estimates of the parameters α_j. It would of course be possible (and desirable) to consider many other specifications as well.

C. *Investment Equations*

As we have just seen, the q theory can be used to formulate equations for the estimation of adjustment cost parameters with many capital goods as well as with a single capital good. The parallel between the two cases is incomplete, however, because while (12) can be inverted to give the rate of investment as a function of Q, this is not possible in the many-good case. That is, the value of Q does not, in general, uniquely determine either the *rate* of investment in value terms, pI/pK, or the *level* of investment, pI. This conclusion is obvious from (18), since it is perfectly clear that the left-hand side is not, in general, equal to a monotonic function of pI/pK, as would be required for the derivation of the investment rate as a function of Q.

To see this negative conclusion in a different way, let us suppose that such an investment equation *can* be written, and ask what restrictions on the model are implied thereby. It is convenient to start with the case where adjustment costs take the form of wasted capital (i.e., $\Gamma \equiv 0$), since it is easier and more complete results are available. I hypothesize that

$$(20) \qquad pI/pK = h(Q),$$

where $h' > 0$, that is, the investment rate depends only on Q. Using (18), this is equivalent to

$$(21) \qquad pI/pK = h\big(\Sigma_j \phi_j'(I_j/K_j)\omega_j\big)$$

where I define $\phi_j(I_j/K_j) = \Phi_j(I_j/K_j, 1)$.

In general, any given total investment rate pI/pK can be the result of widely different levels of investment I_j in individual capital goods, depending on output prices, capital goods prices, variable factor prices, and initial capital stock. If (21) is to obtain in all of these situations, then it should hold identically in I, K, and p. Thus, differentiating with respect to I_j, we have (after cancellation of p_j/pK)

$$(22) \qquad 1 = h'(\cdot)\phi_j''(\cdot), \text{ all } j.$$

This implies that $\phi_j''(\cdot) = \phi_{j'}''(\cdot)$ for all j, j' and for all $I_j, I_{j'}$, which means that $\phi_j''(\cdot) = \alpha$, a constant independent of j, for all j. Using the fact that $\phi_j(0) = 0$ and $\phi_j(\cdot) \geq 0$, this condition can be integrated to yield

$$(23) \qquad \Phi_j(I_j/K_j) = \frac{\alpha}{2}(I_j/K_j)^2, \text{ for all } j;$$

or

$$(23') \qquad \Phi_j(I_j, K_j) = \alpha I_j^2/2K_j, \text{ all } j.$$

In other words, it is possible to write an investment equation like (20), such that the rate of total investment depends uniquely on Q, only if the adjustment cost functions are quadratic and identical. By the same token, it easily follows from (23) and (18) that $Q = \alpha pI/pK$ in this case, so that the identical quadratic specification is sufficient, as well as necessary, for an investment equation like (20) to hold. Note that this implies that the function h is just the adjustment cost parameter, α.

This demonstration can be repeated, essentially intact, for the case where *total* investment pI is expressed as a function $h(Q \cdot pK)$, Q being corrected by pK in order to provide the necessary homogeneity property. Beginning, then, with the analogue of (21), namely,

$$(21') \qquad pI = h\big(\Sigma_j \Phi_{I_j} p_j K_j\big),$$

one derives (22) just as before. Thus:

PROPOSITION 1: *Suppose adjustment costs take the form of wasted capital goods. Then the total investment rate (or total investment) can be expressed as a function of Q (or QpK) if and only if each adjustment cost function is identical and quadratic. In addition, the total investment rate (or total investment) must be linear in Q (or in QpK).*

Now let us consider the case where adjustment costs take the form of lost output. We again suppose (20) holds for all values of I, K, and p, and use (18) to express this relation as

$$(24) \qquad pI/pK - h\big(\Sigma_j p_0 \Gamma_{I_j}\omega_j/p_j\big) = 0.$$

First, differentiating (24) with respect to p_j, yields

$$(25) \qquad p_{j'}I_{j'} = \big[h(\hat{Q}) - h'(\hat{Q})\hat{Q}\big]\, p_{j'}K_{j'}.$$

Summing over j' and using (20), we find that

$$(26) \qquad h(Q) = h(Q) - h'(Q)Q,$$

which cannot hold, given $h' > 0$. Thus, (20) cannot possibly be valid unless all capital good prices are fixed. To draw out the remaining implications of (20), I now make this assumption:

Then (20) *holds if and only if* (24) *holds identically in* (I, K). *By the implicit function theorem, this can occur only if the derivative of*

$pI/pK - h(\Sigma_j p_0(\partial\Gamma/\partial I_j)K_j\omega_j/p_j)$ with respect to every I_j and K_j, is zero. For if this condition is not met, it would be possible to solve for, say, I_1 in terms of $(I_2,\dots,I_m,K)$, which is to say that (24) *would be valid only for some unique value of I_1. Clearly this condition that all $2m$ derivatives of the left-hand side of* (24) *be zero will not, in general, be met.*

The economic and mathematical content of this restriction are given in

PROPOSITION 2: *Suppose adjustment costs take the form of wasted output. Then the total investment rate can be written as a function of Q if and only if* (a) *the relative prices of all capital goods are fixed, and* (b) *the weighted average marginal adjustment cost per dollar's worth of investment in each capital good (where the weights are each good's value share in the initial capital stock) must depend only on the rate of total investment independently of the composition of investment and of the initial capital stock. Condition* (b) *holds if and only if $\Gamma(I, K)$ satisfies*

$$(27)\quad \Sigma_j p_0 \Gamma_{I_j I_{j'}} \omega_j / p_j p_{j'}$$

$$= \frac{\Sigma_j \dfrac{p_0 \Gamma_{I_j} \omega_j}{p_j pK} - \dfrac{p_0 \Gamma_{I_{j''}}}{p_{j''} pK} - \Sigma_j \dfrac{p_0 \Gamma_{I_j K_{j''}} \omega_j}{p_j p_{j''}}}{pI/pK}$$

for all $j', j'' = 1,\dots,m$ and for all (I, K).

The verbal statement of condition (b) in this result follows directly from the interpretation of Q given after equation (18). The formal statement of (b) in equation (27) follows from differentiation of (24) with respect to each I_j, K_j, setting every derivative equal to zero and eliminating h'.

From its verbal statement alone it is evident that not all adjustment cost functions meet condition (b). Similarly, (27) is a system of $2m-1$ equations, none of which would necessarily be satisfied by a general linear homogeneous adjustment cost function. Indeed, the natural question to ask is whether *any* adjustment cost function can satisfy condition (b). Only for the class of additively separable functions have I been able to discover a definitive answer to this question:

COROLLARY: *Suppose that adjustment costs take the form of lost output, that relative capital goods prices are fixed, and that the adjustment cost function is additively separable in I_j, K_j pairs: i.e., $\Gamma(I, K) = \Sigma_j \Gamma_j(I_j, K_j)$. Then condition* (b) *of Proposition* 2 *is met if and only if each Γ_j is a quadratic in I_j/K_j such that the schedule $p_0\Gamma_j/p_j$ of adjustment costs per dollar's worth of investment is identical for all capital goods j, and $h(Q)$ is linear. Formally,* (b) *holds if and only if there is some constant $\alpha > 0$ such that*

$$(28a)\quad \Gamma_j(I_j, K_j) = (\alpha p_j / 2p_0)(I_j/K_j)^2$$

$$(28b)\quad h(Q) = Q/\alpha.$$

The proof of this result is easily patterned after the demonstration of Proposition 1, once it is realized that for fixed p_0 and p, the additively separable $\Gamma = \Sigma_j \Gamma_j$ is formally equivalent to the additively separable Φ_j functions considered in the "wasted capital" case. (It is easily checked that (28a) and (28b) satisfy (27), and conversely that (27) and additive separability imply (28a) and (28b).)

Thus there is definitely one case, albeit a quite special one, in which the total investment rate will be uniquely determined by tax-adjusted Q. Other similar special cases could presumably arise for Γ functions that are not additively separable. Any such Γ must, however, satisfy the $2m-1$ restrictions (27). The unambiguous conclusion must be, as stated earlier, that total investment in many capital goods is not, in general, uniquely determined by a q-type variable.

REFERENCES

Baily, Martin Neil, "Productivity and the Services of Capital and Labor," *Brookings Papers*, *1*:1981, 1–50.

Hayashi, Fumio, "Tobin's Marginal q and Average q: A Neoclassical Interpretation," *Econometrica*, January 1982, *50*, 213–24.

Malkiel, Burton G., von Furstenberg, George M. and Watson, Harry S., "Expectations, Tobin's *q*, and Industry Investment," *Journal of Finance*, May 1979, *34*, 549–61.

Salinger, Michael A. and Summers, Lawrence H., "Tax Reform and Corporate Investment: A Micro-econometric Simulation Study," Working Paper No. 757, National Bureau of Economic Research, 1981.

Summers, Lawrence H., "Taxation and Corporate Investment: A *q*-Theory Approach," *Brookings Papers*, *1*:1981, 67–127.

Tobin, James and Brainard, William C., "Asset Markets and the Cost of Capital," in Bela Belassa and Richard Nelson, eds., *Economic Progress, Private Values and Public Policy*, Amsterdam: North-Holland, 1977, 235–62.

von Furstenberg, George M., "Corporate Investment: Does Market Valuation Matter in the Aggregate?," *Brookings Papers*, *2*:1977, 347–97.

______**, Malkiel, Burton G. and Watson, Harry S.,** "The Distribution of Investment Between Industries: A Microeconomic Application of the '*q*' Ratio," in G. M. von Furstenberg, ed., *Capital, Efficiency, and Growth*, Cambridge: Ballinger, 1980, 395–460.

Yoshikawa, Hiroshi, "On the '*q*' Theory of Investment," *American Economic Review*, September 1980, *70*, 739–43.

Advertising and the Price and Quality of Optometric Services

By JOHN E. KWOKA, JR.*

Restrictions on price advertising have been a long-standing feature of most markets for professional services. The prevailing economic view of these restrictions is that they constitute restraints on the provision of information that maintain excessive prices for such services. Much of the evidence for this view, however, comes from commodities which are standardized and assessable by consumers. By contrast, most professions claim that their services are neither, and that advertising may mislead consumers and undermine the provision of high quality. Market failure may be a danger under such circumstances.

This note examines advertising of professional services, and of optometry in particular. In Section I, I elaborate on the distinctive characteristics of professional services and summarize relevant theory and empirical evidence. Section II analyzes new, detailed data on the effect of advertising on the price and quality of optometric services. The empirical results are consistent with improved market efficiency from advertising, in that prices fall without the erosion of quality sometimes feared.

I. Advertising and Professional Services: The Issues

Many professions contend that the markets for their services are characterized by extremely asymmetric information between buyers and sellers, in conjunction with what we may call endogenous heterogeneous quality. The asymmetry is said to arise since consumers cannot assess, or fully assess, the quality of the services they receive. Services are often complex and inherently judgmental (for example, medical checkups) with effects that are long delayed (title searches), or stochastically indeterminate (prescriptions for ailments). The feedback effects on which the market usually depends are impeded, if present at all. In addition, the quality of the service is not predetermined, but instead the provider can choose or vary the quality supplied. Thus, there is no simple objective measure of an adequate title search or medical checkup, but the services provide to varying degrees for contingencies. The determination of appropriate thoroughness is part of the professional's judgement, that is, quality is endogenously determined.

In such circumstances, professional advertising is alleged to hasten or ensure market problems. Although buyers have difficulty in assessing quality, price is easily understood and price advertising is effective in shifting demand. To accommodate their increased volume of business, advertisers are said to lower quality without market recognition, or at least without full market recognition. This is supposedly accomplished by reducing the crucial input, namely, the professional's own time per customer. Those providers who prefer to offer high-quality service must either suffer a progressive decline of their customer base, or, similarly, lower quality. Thus advertising has increased rewards to low-quality professionals and exacerbated pressures on high-quality providers.

Theory and empirical evidence of varying degrees of relevance to this scenario do exist. For example, for homogeneous or predetermined heterogeneous commodities,[1] theory predicts that advertising may lower price by raising demand elasticity for individual sellers and/or by permitting fuller realization of production scale economies. Such price effects have been demonstrated for

---

*Associate Professor of Economics, George Washington University, Washington, D.C. 20052. Helpful discussions and comments on earlier versions of this paper are gratefully acknowledged from R. Bond, B. Boulier, J. Case, A. Daughety, A. Schwartz, and especially L. Benham.

[1] Predetermined heterogeneous commodities are those for which sellers cannot readily vary quality, although various qualities exist in the market.

eyeglasses by Lee Benham (1972), prescription drugs by John Cady (1976), and retail gasoline by Alex Maurizi (1972) and Howard Marvel (1976). The relationship between advertising and product quality has also been considered, again for predetermined heterogeneous commodities. Philip Nelson (1974; 1978) has argued that even for experience goods, high-quality sellers have more to gain from advertising, via repeat purchases, than low-quality sellers. Yet Richard Schmalensee (1978) identifies circumstances under which low-quality sellers advertise more. In any event, neither author considers the simultaneous role of price, or the possibility that quality is a choice variable of the seller.

The effect of asymmetric information on quality (and price) between buyers and sellers was originally considered by George Akerlof (1970). His used car consumers know market average quality only, and his sellers offer products of predetermined heterogeneity. The inability of high-quality suppliers to be adequately compensated in the market leads to a quality erosion ("lemons") process. More recently, Hayne Leland (1979) has modeled the potentially beneficial role of quality standards in such circumstances, and Eric Bond (1982) has produced evidence that the used truck market has escaped the lemons result.

Thus while advertising may generate lower prices, the effects on quality may be more problematic. Indeed, with endogenous quality for professional services (in contrast to Akerlof's used cars), complete and immediate quality erosion may be prevented only by whatever information consumers may bring to bear, plus the professional's own determination to produce quality.[2] Clearly, of course, some amount of information does exist in the market. Most services produce a subset of immediately measurable effects, and consumers learn more from repeat purchases, second opinions, and others' previous experiences. For "contact" services like medical checkups, where consumers are necessarily present, inputs (such as time, capital equipment) can be monitored. Also, some professionals may develop ways of signalling their quality to potential customers, for example, by adopting nonadvertising modes of operation.

On the production side, advertising may produce benefits in a variety of ways. The above has noted that advertising may permit the fuller realization of production scale economies with likely lower prices. An analogous phenomenon may be at work regarding quality. Low volume may be associated with the deterioration of professional skills and reduced incentive to improve skills. And if there are economies of scale with respect to quality control (say, brand names), larger volume may not only improve information about quality, but quality itself.

Thus, a variety of considerations works to sustain and perhaps even improve quality in the presence of advertising. The models predicting quality erosion may be too simple, and empirical evidence has been sparse. The next section partially remedies this latter problem.

II. Advertising and Optometric Services: The Evidence

Tests of the price and quality effects of advertising require data on at least three categories of professionals: those operating in restrictive environments (nonadvertisers of necessity), nonadvertisers in advertising environments (nonadvertisers by choice) and advertisers, obviously in advertising environments as well. Such data are available from a substantial market experiment in the optometry profession conducted by the Federal Trade Commission (Ronald Bond et al., 1980). Optometry has often been studied,[3] since it, virtually alone among the professions, has long had a rich variety of advertising environments. The present data constitute improvements over those used in most previous studies in three respects.

First, cities are classified on the basis of actual advertising rather than restrictions in laws or codes of ethics.[4] Because instances were found where advertising was either greater or less than permitted, formal restrictions are unreliable guides to actual practice. Ultimately, three cities were selected which prohibited all advertising of eye examinations (and eyeglasses as well), and also prohibited were commercial practices such as an optometrist working for an optical firm. All optometrists in these "restrictive" markets were, of course, nonadvertisers. In the four "advertising" markets, price and nonprice advertising of both examinations and eyeglasses were permitted, as well as commercial practice.[5]

---

[2] A model along these latter lines is developed in my working paper, where I demonstrate conditions under which quality erosion does not occur even with extreme informational asymmetry.

[3] See particularly Roger Feldman and James Begun (1978; 1980).

Second, in advertising markets, nonadvertisers are distinguished from advertisers, thereby permitting analysis of the quality erosion hypothesis. In fact, four different types of optometrists are identified, corresponding approximately to the magnitude of their advertising effort. The term *NONE* denotes nonadvertising practitioners; their names appear only in listings of optometrists in the Yellow Pages, and they practice in thoroughly "professional-looking" offices. The term *STORE* denotes on-site advertisers who do not advertise in the Yellow Pages or newspapers, but instead have storefronts with prominent signs or displays. The variable *SMED* denotes optometrists, who are small-firm media advertisers, those affiliated with small local firms (usually optical firms) which advertise in the Yellow Pages and/or newspapers, and *LMED* represents large-chain advertisers, optometrists affiliated with large regional or national optical firms, and heavy advertisers in the media.

Third, the methodology produced primary data of enormous detail and quality, since the data were generated by examinations actually obtained from a stratified sample of optometrists in the selected cities. The seven subjects with similar visual conditions were trained at two schools of optometry with regard to the components of an optometric examination. They obtained examinations and recorded price, time spent, and details on the various tests and procedures performed. The present data set consists of 147 observations.

Table 1 reports regressions which test the effects of advertising on price and one measure of the quality of optometric services. In all regressions, the comparison group (the constant term) consists of all optometrists in restrictive markets, that is, they are nonadvertisers by definition. The variables denoted *NONE*, *STORE*, *SMED*, and *LMED* represent the categories of practitioners previously defined in advertising markets. In addition, following conventional practice, variables are included to control for the number of optometrists per capita in the market (*ODPC*), to represent the intensity of monopolistic competition;[6] income per capita of the population (*YPC*), reflecting different demand conditions; as well as a set of six dummy variables for the subjects.

The first three regressions in Table 1 explore price. Those lettered (a) and (b) are ordinary least squares, and differ only in the inclusion of subject dummies in (b). Both suggest that the presence of advertising causes substantial and significant declines in the prices of eye examinations offered by all types of optometrists. Large- and small-firm advertisers (*LMED* and *SMED*, respectively) appear to charge \$11 to \$12 less than optometrists in restrictive markets, a finding consistent with evidence previously cited. The negative coefficient on the *NONE* variable confirms that even nonadvertisers in advertising markets are forced to lower prices to compete. Yet their reduction is much less than that by advertisers themselves.[7] The addition of six subject dummies in price

---

[4] Specifically, the telephone book Yellow Pages were checked, and newspapers were scanned for two to four months to ascertain the type and amount of advertising. Further details on methodology can be found in Bond et al.

[5] Restrictions on advertising and commercial practice tend to occur simultaneously, and present data reflect that coincidence. To that extent, estimates below may represent an upper bound on advertising alone.

[6] There is reason to believe that the number of optometrists per capita may be greater where advertising exists (see Lee Benham and Alexandra Benham, 1975). If so, present estimates understate the full effect of advertising.

[7] In their study of eyeglass prices, Benham and Benham find similar effects of advertising on the prices charged by nonadvertisers.

TABLE 1—REGRESSION RESULTS FOR PRICE AND TIME-QUALITY OF SERVICE[a]

Dependent Variable	Constant	*NONE*	*STORE*	*SMED*	*LMED*	*ODPC*	*YPC*	R^2	F
(a) *Price*	28.92	−2.59	−7.56	−11.13	−12.35	−9.88	8.24	.56	29.4
	(7.79)	(1.94)	(3.16)	(7.48)	(8.68)	(6.54)	(1.10)		
(b) *Price*[b]	29.18	−2.56	−7.44	−11.03	−12.06	−9.61	8.04	.57	14.6
	(7.54)	(1.88)	(3.00)	(7.30)	(8.27)	(6.14)	(1.05)		
(c) *Price*[c]	30.63	−1.40	−6.13	−9.88	−11.19	−8.52	1.44	.64	41.5
	(10.41)	(.98)	(2.15)	(5.90)	(8.43)	(5.30)	(.24)		
(d) *Time*	40.07	7.68	−4.61	−7.01	−8.01	2.14	−34.02	.29	9.3
	(3.95)	(2.11)	(.71)	(1.72)	(2.06)	(.52)	(1.67)		
(e) *Time*[b]	49.53	7.55	−4.23	−6.92	−8.09	2.49	−32.33	.40	7.4
	(5.05)	(2.18)	(.67)	(1.81)	(2.19)	(.63)	(1.67)		
(f) *Time*[b,c]	41.23	11.71	−0.87	−3.14	−4.89	9.51	−27.84	.52	12.0
	(6.48)	(2.79)	(.18)	(.80)	(1.29)	(2.05)	(2.02)		

[a] The t-statistics are shown in parentheses.

[b] Subject dummies are included in regression.

[c] *GLS* regression. The constant term is a variable representing between-group differences in sample variance. The statistic F is calculated by comparing the explanatory power of a regression where all the coefficients except the constant are constrained to zero; R^2 corresponds to that F. See George Judge et al. (1980, pp. 254–56).

regression (b) does not affect the magnitude or significance of any result in equation (a). An F-test on the increase in regression sum of squares from the subject dummies yield $F(6,134)=.45$, far short of statistical significance. Subject-specific variation is not important in the price regressions.

A correction which is of some importance, however, is for heteroskedasticity among the five types of optometrists.[8] Using generalized least squares in equation (c), we obtain efficient parameter estimates in most cases close to the previous values. As before, all optometrists in advertising markets lower their prices, from $1.40 by nonadvertisers to over $11 by large-firm advertisers. The statistical significance of the nonadvertisers' price reduction has diminished, although the difference between *NONE* and *LMED* remains high significant ($t=13.9$). These differences can be compared to a restrictive market mean price of $25.17, evaluated at the means of *ODPC* and *YPC*. The latter, it should be noted, have the expected signs, although per capita income is not statistically significant.

Regression equations (d), (e), and (f) in Table 1 perform the same tests on the measure of quality: time (in minutes) spent in the examination. In the absence of a standard definition of quality in optometry, "time" has the advantages of being unbiased, observable, and correlated with output measures of quality.[9] And, as previously noted, in many circumstances it is the crucial input to quality, supposedly reduced as the result of advertising. In addition to the type of optometrist and the other control variables in equation (d), dummy variables for subjects are added in equation (e). In this case, their impact is more pronounced, and collectively they are statistically significant with $F(6,134)=4.16$. The effects of subjects on time-quality is probably due to slight differences in their visual conditions, requiring different procedures, and/or to individual differences in their mannerisms during examinations. I shall focus on equations which control for such extraneous factors.

In addition, heteroskedasticity among the types of optometrists requires application of generalized least squares (*GLS*).[10] As reported in equation (f), the possibility that advertisers give shorter examinations is supported. The coefficient on *LMED* implies

---

[8] A likelihood-ratio test for equality of the price regression residuals among the five types of optometrists yields a test statistic $\lambda=23.8$. The 5 percent region with four degrees of freedom is defined by $\lambda=9.5$. See Jan Kmenta (1971, p. 268). The standard deviations of the computed residuals are used to normalize the data.

[9] In the Bond et. al. study, consulting optometrists developed an "index of thoroughness" by weighting the various tests and procedures. Its simple correlation with the time input is .71.

[10] The test statistic for equality of variances is $\lambda=47.1$. See fn. 8.

TABLE 2—*GLS* REGRESSIONS FOR PRICE, TIME-QUALITY HELD CONSTANT

Constant	*Time*	*Time* × *Adv*	*NONE*	*STORE*	*SMED*	*LMED*	*ODPC*	*YPC*	R^2	*F*
(a) 24.16	.133		−2.83	−6.06	−9.39	−10.52	−9.19	9.23	.70	46.7
(7.87)	(4.71)		(2.05)	(1.88)	(5.64)	(7.99)	(5.82)	(1.55)		
(b) 25.47	.064	.095	−5.48	−8.40	−11.65	−12.75	−9.14	10.21	.71	41.5
(8.02)	(1.17)	(1.51)	(2.46)	(2.36)	(5.21)	(6.44)	(5.81)	(1.71)		

Note: See footnotes to Table 1.

that large-firm optometrists spend about 5 minutes less in a typical examination than practitioners in restrictive markets. This difference is statistically significant at 10 percent in a one-tail test, and bears out one portion of the argument made by optometrists and other professionals concerning the effects of advertising. But a crucial part of their story, that concerning the fate of nonadvertisers, is not confirmed. Nonadvertisers in advertising markets give higher quality examinations than their counterparts in restrictive markets. The coefficient on *NONE* indicates their exams are over 11 minutes *longer* than the restrictive market average of 24 minutes.

This conclusion stands in sharp contrast to the argument that advertising induces an erosion of quality by all practitioners in the market. These data reveal that nonadvertisers-by-choice actually provide service of superior quality to nonadvertisers in restrictive markets. This result may be due to (*i*) simple disaggregation of all practitioners in restrictive markets, such that those who provide above-average quality adopt a nonadvertising mode; (*ii*) lower opportunity cost of time to nonadvertisers in advertising markets, who lose customers to advertisers (unless they take countermeasures); or (*iii*) greater sensitivity of consumers in advertising markets to quality differences, with consequently stronger feedbacks for high quality. In any case, such nonadvertisers constitute a sizeable fraction of providers in advertising markets. Weighting the time differences in regression (f) by the frequency of each kind of optometrist found,[11] the mean time spent in advertising markets is actually *longer* than in restrictive markets.[12] This is altogether inconsistent with the view that high-quality service is endangered by advertising.

One final matter concerns the overall relation between price and time-quality. If such an association exists, it must be taken into account in estimating the effect of removing advertising restriction, since only constant-quality prices will have significance. Table 2 reports the results of *GLS* price regressions analogous to equation (c) in Table 1, with the addition of variables to hold quality constant. Equation (a) of Table 2 demonstrates the strong positive association throughout the data between price and time-quality ($t = 4.71$). The significantly different intercepts for the four categories of practitioners in advertising markets, even after controlling for time, are price differences associated with other dimensions of quality (for example, waiting time, location), and might be due to residual imperfections in consumer information.

A variation of interest is to add a slope dummy on the time variable permitting a different relationship between price and quality in advertising markets. In equation (b) of Table 2, the variable *Time* × *Adv* has a sizeable positive coefficient, and $t = 1.51$, significant at about 13 percent. It would appear that price of *additional* examination time may be higher in advertising environments, even while total price is less. This would be consistent with practitioners' greater ability to differentiate their services, and seemingly to get compensated accordingly, in advertising markets. By contrast, restrictive markets cause a leveling of prices despite the existence of quality diversity.

---

[11] These weights are obtained from the proportions of various types of optometrists in the Yellow Pages, and in field work. Thus, the price comparisons are of offer prices, not transactions prices in the market. The larger share of advertisers would make the price difference with restrictive markets greater yet.

[12] The difference is 5.3 minutes. Recall that these numbers represent quality offered, and do not reflect the likely larger share of chain firms. In addition, it is interesting to note that the variance of time spent in restrictive markets is 133.9 (standard deviation of 11.6 minutes) compared to 118.8 (standard deviation of 10.9 minutes) in advertising markets. The difference, however, is not significant, with $F(45,97) = 1.13$.

From the results in equation (a), we apply weights representing the proportions of each type of practitioner to the estimates of constant-quality price differences in advertising markets. The mean difference is $5.23. Compared to an average price in restrictive markets of $25.44,[13] this implies that the removal of advertising restrictions would cause constant-quality prices to decline over 20 percent. And to reiterate, this occurs without a decline in overall market quality.

III. Summary

This paper has explored a market process that, many professions argue, may be precipitated by advertising. In particular, the claim that advertising lowers price and quality, forcing others to do so, is tested with data from the optometry profession. The evidence reveals that advertisers' prices and quality are indeed lower, and while nonadvertisers' prices fall, their quality actually is greater. Furthermore, nonadvertisers remain in sufficient numbers in the market so that average market quality is not lower, but indeed greater. These results are inconsistent with key elements of the professions' argument against advertising, but instead imply considerable social benefits from loosening restrictions as shown on the practice of optometry.

---

[13] This amount differs slightly from that derived before because of the additional control for quality in the present regression.

REFERENCES

Akerlof, George, "The Market for 'Lemons': Quality Uncertainty and the Market Mechanism," *Quarterly Journal of Economics*, August 1970, *84*, 488–500.

Benham, Lee, "The Effect of Advertising on the Price of Eyeglasses," *Journal of Law and Economics*, October 1972, *15*, 337–52.

______ **and Benham, Alexandra,** "Regulating Through the Professions: A Perspective on Information Control," *Journal of Law and Economics*, October 1975, *18*, 421–47.

Bond, Eric, "A Direct Test of the 'Lemons' Model: The Market for Used Pickup Trucks," *American Economic Review*, September 1982, *72*, 836–40.

Bond, Ronald et al., *Effects of Restrictions on Advertising and Commercial Practice in the Professions*, Washington: Federal Trade Commission, 1980.

Cady, John, "An Estimate of the Price Effects of Restrictions on Drug Price Advertising," *Economic Inquiry*, December 1976, *14*, 493–510.

Feldman, Roger and Begun, James, "The Effects of Advertising Restrictions—Lessons from Optometry," *Journal of Human Resources*, Supp., 1978, *13*, 247–62.

______ **and** ______, "Does Advertising Reduce the Mean and Variance of Prices?," *Economic Inquiry*, July 1980, *18*, 479–84.

Judge, George et al., *The Theory and Practice of Econometrics*, New York: Wiley & Sons, 1980.

Kmenta, Jan, *Elements of Econometrics*, New York: Macmillan, 1971.

Kwoka, John, "Advertising and the Price and Quality of Optometric Service," Department of Economics Discussion Paper 8209, George Washington University, 1982.

Leland, Hayne, "Quacks, Lemons, and Licensing: A Theory of Minimum Quality Standards," *Journal of Political Economy*, December 1979, *87*, 1328–46.

Marvel, Howard, "The Economics of Information and Retail Gasoline Price Behavior: An Empirical Analysis," *Journal of Political Economy*, October 1976, *84*, 1033–66.

Maurizi, Alex, "The Effects of Laws Against Price Advertising: The Case of Retail Gasoline," *Western Economic Journal*, September 1972, *10*, 321–29.

Nelson, Philip, "Advertising as Information," *Journal of Political Economy*, July/August 1974, *81*, 729–54.

______, "Advertising as Information Once More," in David Tuerck, ed., *Issues in Advertising*, Washington: American Enterprise Institute, 1978, 133–60.

Schmalensee, Richard, "A Model of Advertising and Product Quality," *Journal of Political Economy*, June 1978, *86*, 485–503.

A Critical Appraisal of McKinnon's World Money Supply Hypothesis

By HENRY N. GOLDSTEIN AND STEPHEN E. HAYNES*

In 1978, Arnold Harberger pointed to the similarity of inflation rates across sixteen industrial countries during each of the three periods: 1952–67; 1967–72; 1972–78. Noting the relatively narrow range of average inflation rates for the middle twelve countries in each period, Harberger conjectured that the driving force behind this shared inflationary experience was *the common impact of the world money supply under a regime of relatively fixed exchange rates*. Although the major currencies have been nominally floating against one another since early 1973, he nonetheless contended that the concept of a world money supply remained a fruitful notion:

> Even though we have had nearly five years of floating rates, it is my impression that the monetary authorities of the major countries have behaved with far less independence than the theorists of floating rates are prone to assume. Consequently, in terms of the way the international monetary system has worked since 1973, I would characterize it as a hybrid, functioning if anything more like a fixed-exchange rate system than like a textbook case of floating rates. [p. 511]

Much the same evaluation has recently been made by Ronald McKinnon (1982). According to McKinnon, a combination of intense intermittent speculative pressure, heavy official intervention, and a "perfect" world capital market has meant that, even under floating exchange rates, the world money supply has had a significant impact on inflation and output in the United States. McKinnon concludes that the Federal Reserve should adjust its money-growth policies in the light of money supply behavior in other leading countries. This behavior, he argues, moves in close correspondence with the exchange rate between their currencies and the dollar. Thus, he writes: "The doctrine of domestic monetarism, where the Federal Reserve System keys on some purely American monetary aggregate such as $M1$ and $M2$ and ignores the foreign exchanges, is increasingly inefficient for preventing global inflation or deflation—and for stabilizing American income and prices" (p. 332).

We dispute McKinnon's interpretation of the behavior of exchange rates and national money supplies since the early 1970's. This paper includes (Section I), a brief restatement of McKinnon's central point; (Section II), several criticisms of his analysis and an alternative interpretation of events under managed floating; and (Section III), some evidence based on "St. Louis-style" regressions suggesting that quarterly money supply growth in other industrial countries has had only a small impact on U.S. prices and activity levels.

I. McKinnon's Model

McKinnon's model has two countries, the United States and *ROW* (rest of the world), with the dollar and the *rowa* as national currencies. The spot and forward dollar prices of the *rowa* are denoted by S and F, and interest rates on dollar-denominated and *rowa*-denominated assets by i and i^*. Now, consider an initial equilibrium where, for simplicity, s, the expected near-term change in the spot price of the *rowa* [$s = E(dS/dt) = (F - S)/S$] is zero, the latter equality holding because F is assumed equal to the expected value of S at the relevant future date. Satisfaction of McKinnon's equation

---

*Professor and Associate Professor of Economics, University of Oregon, Eugene, OR 97403. For helpful comments, we thank Peter B. Clark, George G. Kaufman, Jay H. Levin, J. Carter Murphy, Joe A. Stone, and Steven Strongin.

$$i - i^* = s \text{ (Fisher Open Condition)} \tag{6}$$

then implies that $i_0 = i_0^*$, where the subscript denotes time zero. Now let there be a sharp exogenous rise in s. According to McKinnon, this development triggers currency substitution through two channels. Channel One comprises direct shifts of working balances from dollars into *rowa* by large commercial banks and nonfinancial multinationals. Channel Two involves indirect currency substitution which, McKinnon argues, is likely to be much more important than direct substitution. It is the discovery and elucidation of this Channel Two process that represents the key contribution of the McKinnon article.

The Channel Two process works as follows. As s increases exogenously, incipient pressure by international bond arbitragers increases i and decreases i^* until the spread between the interest rates $(i_1 - i_1^*)$ equals the new value of s. Simultaneously, the forward discount on the dollar takes a value equal to s. At the new interest rates, transactors in each country find themselves off their respective money demand curves. The U.S. transactors now want to reduce their holdings of dollar balances, but the U.S. money supply remains unchanged; similarly, *ROW* transactors want to increase their holdings of *rowa* balances but the *ROW* money supply remains unchanged. The U.S. transactors respond by trying to switch out of cash balances and into dollar-denominated bonds; and *ROW* transactors by trying to switch out of *rowa*-denominated bonds and into *rowa* cash balances. *The result is to push* $(i - i^*)$ *just barely below* s, which induces international bond arbitragers—sensitive to even the slightest expected uncovered yield differential—to switch out of dollar bonds and into *rowa* bonds. Thus, McKinnon concludes "*Massive capital flows can easily be induced even when the interest differential remains 'correctly' aligned to reflect accurately the change in expected exchange depreciation*" (p. 327).

What is the response of the central banks to these capital flows? McKinnon assumes that the *ROW* central bank intervenes to peg the spot exchange rate in the immediate short run. As a result, it creates new high-powered balances in *rowa* on a scale sufficient to increase *ROW* money to the amount demanded at the lower interest rate, i_1^*. Moreover, the *ROW* central bank cannot sterilize the domestic monetary base effects of these speculatively induced capital inflows because of *perfect capital mobility*—a key assumption of McKinnon's model. In the United States, however, such sterilization occurs automatically insofar as the *ROW* bank uses its dollar accruals to buy U.S. government securities rather than increase its dollar deposits at the Federal Reserve. This sterilization means that dollar-denominated transaction balances in the United States exceed the amount demanded. This disequilibrium is presumably resolved by persisting attempts by U.S. transactors to acquire both bonds *and* goods, with resultant lagged expansionary pressures on the world economy. As nominal world income responds, the money demand curves in both the United States and *ROW* shift to the right by amounts sufficient to restore equilibrium in the combined "markets" for the two national money supplies.

Sterilization by the Federal Reserve represents a key policy mistake, according to McKinnon, for it implies that an exogenous shift in exchange rate expectations, as indicated by an increase in the value of the parameter s, leads to the creation of an excess supply of *world money*. In a dynamic context, what the Fed authorities have failed to perceive is that the more rapid growth of foreign money balances (stemming from speculatively induced capital flows out of the dollar) needs to be balanced by a slower growth of U.S. money balances. By focusing only on the growth rate of U.S. money, the Fed has aimed at the wrong target. McKinnon uses his model to explain "...the two explosions in world money supply in 1971–72 and again in 1977–78" (p. 331) when, he argues, s took on large positive values, and "The Great Deflation of 1981–82" (p. 332, Section V) when he claims that there was an exogenous fall in s.

II. Questions about the McKinnon Model and an Alternative View

Several important elements of the McKinnon model do not appear to be plausible and realistic descriptions of the recent managed float.

1) McKinnon assumes that huge amounts of uncovered funds move in response to very small differentials in expected yields. Such sensitivity seems implausible for uncovered capital movements. Though international borrowers and lenders may well "take a view" regarding the probable level of the spot rate in the near future, it seems unlikely that they would expose themselves to large adverse swings in the exchange rate merely to obtain a 10 or 20 basis-point incremental expected return. In brief, for risk-adverse transactors, the uncertainty of this expected incremental return seems likely to outweigh the appeal of its small expected value.[1]

2) To make the Fisher open condition consistent with observed interest differentials of, for example, 2 to 10 percent per annum would require assigning s (the expected *near-term* change in the spot price of *rowa*) a monthly value of 0.16 to 0.83 percent. Such values seem implausibly small given the state of foreign exchange markets during periods of particularly strong one-way pressure.

3) Consider the recent period of strong dollar appreciation. From mid-1980 to mid-1982, the dollar appreciated substantially against almost all other leading currencies. The *DM* was a leading example. As the Bundesbank leaned against the wind during this period, its external reserves, exclusive of gold, fell from $52.5 billion at end-1979 to $41.0 billion at end-July 1982. If the market expected the *DM* to depreciate during this period, as McKinnon assumes, his model requires that short-term *DM* yields exceed short-term dollar yields. But in fact, except for a brief period in mid-1980, rates on dollar assets remained well *above* rates on comparable *DM* assets. Indeed, German domestic interest rates have been substantially and consistently below U.S. domestic interest rates both when the *DM* was strong against the dollar (1977–79) and when it was weak against the dollar (1980–mid-1982). One does not have to search hard for an explanation: the greater expected inflation in the United States than in Germany. Similar bilateral comparisons, contrary to McKinnon's model, could be made between the United States and Switzerland, as well as with Japan and the Netherlands. In other words, consistent with empirical evidence in Bruno Solnik (1982), satisfaction of "the Fisher closed condition" may well have prevented satisfaction of the "Fisher open condition" in a world where exchange rates were not expected to move in strict accordance with relative changes in national price levels.[2]

An alternative assessment of the managed floating rate experience since 1973—presented in somewhat dogmatic fashion as a foil to the McKinnon view—runs as follows:

1) By and large, *ROW* central banks do "take a view" regarding the appropriate level and rate of change of the dollar price for their currencies. If the rate begins to diverge from the level they think appropriate, they will usually intervene. But they will only lean against the wind—they will not prevent market pressure from moving the rate.[3] Accordingly, we disagree with Harberger's perception that since 1973 the hybrid international monetary system has functioned "more like a fixed-exchange rate system than a textbook case of floating rates" (p. 511).

2) The central banks also have short-run interest rate and money supply targets, which they vary in response to changes in domestic demand conditions and exogenous inflationary shocks. By and large, the central banks try to neutralize the effects of their dollar

---

[1] Bruno Solnik (1982) has recently tested whether international interest-rate spreads for the period 1971–1980 can be better explained by expected exchange rate changes or by differences in domestic economic policies. He finds "much more support for a domestic model of interest rate determination" (p. 337). "This result," he notes, "could be explained by international market inefficiencies but more probably [is explained] by the fact that exchange rate variations are so unpredictable that their expectations play a little role in setting differences between national interest rates" (p. 333).

[2] Examination of returns on assets issued in the same political jurisdiction but in different currencies may lead to findings more consistent with the Fisher open condition; for an analysis of interest rate parity using such returns, see Jacob Frenkel and Richard Levich (1975). Even with some independence between corresponding domestic and international interest rates, however, the interest rate relevant to monetary policy and the McKinnon policy prescription is domestic rather than international.

[3] For a similar view citing a number of empirical studies, see Paul Wonnacott (1982, p. 3).

accruals or sales on their domestic monetary base, but if the exchange market pressure is intense they may choose to alter their earlier domestic money-growth and interest rate targets at the same time that they back away from (yet resist) that pressure. Such a response makes good economic sense because the exchange rate trend influences prices and real demand in a way that offsets the effects of the faster or slower monetary growth induced by their intervention. During 1977–78, for example, Germany experienced low inflation relative to her major trading partners together with considerable slack capacity. As the *DM* strengthened, the Bundesbank absorbed substantial amounts of dollars from the exchange market. Its policy of only partially sterilizing, hence allowing the domestic money supply to grow well above the earlier target range, seems sensible *given that faster money growth was accompanied by a sharply rising DM in the foreign exchange market.*[4]

3) Slower monetary growth in the non-U.S. industrial countries during 1980–82 represented a response to cost-push pressures caused by the second oil shock and a strengthening dollar; it was not caused by an inability to offset the domestic-base effects of dollar sales in the exchange market.[5] Indeed, these effects *were* more than fully offset in that money growth remained positive even though the central banks were net sellers of dollars in the exchange markets.

4) We have argued that large Channel Two capital flows are implausible. We suspect that Channel One capital flows are also unimportant. Aside from compensating balances, sophisticated multinational concerns are unlikely to hold significant amounts of transactions balances in any currency if these balances yield either zero or small explicit interest returns. Such transactors might well choose to go long or short in a given currency for speculative reasons, but in so doing they would prefer to acquire an interest-earning asset.[6]

III. Some St. Louis Regressions

Here we present regression evidence which seems consistent with our assessment of the post-1973 international monetary system and inconsistent with that of Harberger and McKinnon.[7] The regressions test whether the

---

[4] As the Bundesbank noted in its 1978 *Report:*

> Both at the beginning of the year [1978] and in the autumn new monetary unrest associated with the temporary weakness of the U.S. dollar and to some extent also with tensions in the European narrower margins arrangement (the "snake"), resulted in a rapid appreciation of the Deutsche Mark in international foreign exchange markets. In order to slow down this upward movement, *which far exceeded the price and costs differential between Germany and other countries*, and at the same time to mitigate the dampening effect of an excessive appreciation on domestic activity, the Bundesbank took considerable amounts of foreign currency out of the market.... In these circumstances the Bundesbank was forced to adhere to a relatively easy domestic monetary policy. Both factors—the massive inflows of funds from abroad and the stance of domestic monetary policy—led to an expansion of the money growth target for 1978. *As long as the trend of business activity remained fairly flat (as in the early part of last year) and as the sharp appreciation of the Deutsche Mark automatically restrained the increase in domestic costs and prices, it was possible to tolerate the faster expansion of the liquid funds in Germany*; in fact, any application of the monetary brakes during that period would probably have had an adverse effect on the economy as a whole.
>
> [p. 1, emphasis added]

[5] On this issue, see the comments by Herman Lehment (1982) and Kumihara Shigehara (1982).

[6] For supporting evidence that the demand for domestic money balances does not depend on foreign rates of interest, see Michael Bordo and Ehsan Choudhri (1982a).

[7] Three pieces of less formal evidence also seem inconsistent with the ideas of Harberger and McKinnon: First, for the ten leading OECD industrial countries (Belgium, Canada, France, Germany, Italy, Japan, Netherlands, Switzerland, United Kingdom, and United States) the percent changes in the *CPI* and in the money supply were compared over two six-year periods—1963–69, and 1976–82. For the earlier period, the percent change in the money supply ranged across countries from 35 to 165 percent, yet the percent change in the *CPI* ranged only from 10 to 40 percent, suggesting little policy independence under fixed exchange rates; for the latter period, however, the two percent changes were similar for each country, ranging from 25 to 180 percent, suggesting substantial policy independence under the float. Second, the dispersion in annual inflation rates for the same ten countries was examined over the two different policy regimes. A marked increase in dispersion after the adoption of floating rates was evident, again suggesting an increase in policy independence under that regime. Third, as a test of the sterilization assumption of McKinnon, we regressed for the OECD countries (other than the United States) the percentage change in the country's money supply on a distributed lag of the percentage change in its international reserves for the 50 quarters ending in mid-1982. Only three coefficients in the nine regressions were significant, and

TABLE 1—REGRESSION ESTIMATES OF U.S. INFLATION: 1973:II–1982:II

Equation	Constant	*MUS*3	*MW*3	*MROW*3	*ENER*	*DUMA*	*D-W*	R^2	$SE \times 10^2$
(a)	−.003	1.091			.035	.042	2.15	.519	.397
	(−.039)	(3.56)			(1.43)	(2.28)			
(b)	.024		.479		.069	.018	1.37	.326	.470
	(.751)		(1.26)		(2.57)	(.892)			
(c)	−.019	1.155		.145	.037	.040	2.17	.531	.406
	(−.636)	(3.61)		(.782)	(1.49)	(2.12)			

Note: The *t*-statistics are shown in parentheses. The dependent variable is the annual percentage change in the *GNP* deflator for the quarter; *MUS*3 is the average annual rate of change in the U.S. money supply *M*1 over the previous 12 quarters; *MW*3 (*MROW*3) is the average annual rate of change in the world (*ROW*) money supply over the previous 12 quarters; *ENER* is the average annual rate of change in an index of energy prices divided by the *GNP* deflator over the current and previous 5 quarters; *DUMA* is a postprice control dummy (equal to one over 1974:II–1974:IV, zero elsewhere). For data source, see fn. 9.

weighted average growth rate of money supplies in industrial countries outside the United States has had a significant influence on either the U.S. rate of inflation or U.S. real output. We follow McKinnon in measuring the growth rate of world money where the world consists of ten industrial countries —the United States, Canada, Japan, the U.K., Germany, France, Italy, the Netherlands, Belgium, and Switzerland. For any given quarter, growth rates in these countries' narrowly defined seasonally adjusted money supplies were computed, weighted by their relative *GNP* levels at midsample, and then summed.[8] A corresponding series for the growth rate of Rest-of-World (*ROW*) money was constructed by omitting the U.S. component from the world money index and rescaling the remaining weights to make them again sum to one.

A. *The U.S. Inflation*

Reduced-form regressions reported by Keith Carlson (1980), Michael Bordo and Ehsan Choudhri (1982b), and others at the Federal Reserve Bank of St. Louis show that quarterly U.S. inflation during the 1970's can be well explained statistically by a regression of the *GNP* deflator on (*i*) lagged U.S. money growth over the past three years; (*ii*) changes in relative energy prices; (*iii*) changes in relative food prices; and (*iv*) dummy variables to capture the effects of wage and price controls. McKinnon's contention that monetary developments in the *ROW* have had powerful effects on the U.S. economy can be tested by using the St. Louis approach. Specifically, suppose that we estimate regressions explaining U.S. inflation for the floating-rate period using moving averages of past U.S. and world money growth as alternative variables—would world money growth prove to be the better explanatory variable, as McKinnon's argument clearly implies? And suppose that we estimate a regression that includes moving averages of *both* U.S. and *ROW* money growth as explanatory variables—would *ROW* money growth prove to be statistically significant and would its inclusion appreciably improve the explanatory power of the equation?

Tables 1 and 2 show the results of these regressions, with the equations in Table 1 covering the floating-rate period, per se, 1973:II–1982:II and those in Table 2, a slightly longer period, 1970:I–1982:II.[9] The

---

each was small in magnitude (below 0.1), evidence consistent with Leroy Laney and Thomas Willett (1982), who argue that increases in international reserves were not the dominant explanation for the global monetary expansion during the 1970's. (For more formal tests of sterilization which attempt to control for possible simultaneity bias, see Maurice Obstfeld, 1982, and the survey by Victor Argy, 1982.)

[8] McKinnon used *GNP* weights for 1970, the midsample year for the two decades of data he presented in his article; we used 1976 weights, as 1976 is the midsample year for most of our regressions. Altering this year does not change the weights markedly.

[9] Money supply data are from International Monetary Fund, *International Financial Statistics*; the U.S. *GNP* deflator and real *GNP* are from the U.S. Department of Commerce; and the energy price data are from the Federal Reserve Bank of Chicago.

TABLE 2—REGRESSION ESTIMATES OF U.S. INFLATION: 1970:I–1982:II

Equation	Constant	*MUS*3	*MW*3	*MROW*3	*ENER*	*DUMC*	*DUMA*	*D-W*	R^2	$SE \times 10^2$
(a)	−.003	1.120			.041	−.010	.032	2.03	.624	.375
	(−0.212)	(4.25)			(1.85)	(−1.80)	(2.87)			
(b)	.005		.699		.077	−.023	.006	1.39	.527	.420
	(.205)		(2.57)		(3.28)	(−2.98)	(.530)			
(c)	−.026	1.136		.208	.046	−0.164	.025	2.10	.644	.369
	(−1.158)	(4.37)		(1.44)	(2.08)	(−2.32)	(2.14)			

Note: See Table 1; *DUMC* is a price control dummy (equal to one over 1971:II–1974:I, zero elsewhere).

TABLE 3—REGRESSION ESTIMATES OF U.S. REAL GROWTH: 1970:I–1982:II

Equation	Constant	*MUS-dev*	*MW-dev*	*MROW-dev*	*ENER*	*D-W*	R^2	$SE \times 10^2$
(a)	.012	.130			−.038	1.97	.370	.985
	(4.48)	(2.19)			(−2.72)			
(b)	.012		.144		−.032	2.02	.365	.992
	(4.41)		(2.11)		(−2.13)			
(c)	.012	.124		.049	−.033	2.08	.389	.987
	(4.35)	(2.09)		(1.06)	(−2.20)			

Note: The *t*-statistics are shown in parentheses. The dependent variable is the annualized percentage growth rate in U.S. real *GNP*, seasonally adjusted; *MUS-dev* is the 2-quarter growth rate for *M*1 over the current and past quarter less the 12-quarter moving average of *M*1; *MW-dev* (*MROW-dev*) is a similar variable for world (*ROW*) money; *ENER* is the average annual rate of change in an index of energy prices divided by the *GNP* deflator over the current and previous five quarters. (The variables *DUMC* and *DUMA*, defined earlier, were also included in the regressions, but their coefficients were insignificant.) For data source, see fn. 9.

results for both periods are quite similar. Comparing equations (a) and (b), we see that U.S. money growth is highly significant in both periods, whereas world money growth is only significant for the longer period. For each period, the explanatory power of the equations using world money is markedly lower than the equation using U.S. money. Moreover, the estimated coefficients for world money growth are well below unity in both cases, whereas the coefficients on U.S. money growth—1.09 for the shorter period and 1.12 for the longer period—are not significantly different from unity.

Equation (c), which includes both *ROW* and U.S. money, is also quite similar for both periods. The growth rate in *ROW* money is statistically insignificant whereas the growth rate for U.S. money is highly significant and its coefficient is little changed in numerical size from that estimated in equation (a). Overall, the addition of *ROW* money adds virtually nothing to the ability of this simplified St. Louis equation to explain U.S. inflation. Reestimation after expressing the money variables in explicit distributed lags rather than moving averages does not alter this conclusion, which suggests that the evidence in Tables 1 and 2 does not merely reflect secular or trend effects.

B. *The U.S. Real Growth*

As Milton Friedman and Anna Schwartz (1963), William Poole (1975), and others have shown, fluctuations in real U.S. economic activity are linked to significant changes in U.S. *M*1 growth from its recent trend. Is it possible that deviations from trend in world money growth have an even stronger correlation with the growth rate of real U.S. *GNP*? And if we were to put the U.S. and *ROW* monetary deviation variables into the same regression, would both prove to be significant, and would the regression have significantly greater explanatory power than the equation without the *ROW* money-growth variable? McKinnon's analysis suggests positive answers to these questions.

To test this implication of his analysis, we estimated the regressions shown in Table 3. Equations (a) and (b), which include the U.S.

and world monetary deviation variables separately, show virtually no difference in their ability to account for fluctuations in U.S. real *GNP*. On the other hand, in equation (c), the U.S. monetary variable remains significant and with a much larger coefficient than the *ROW* monetary variable; and, indeed, *MROW-dev* is not statistically significant. As with the earlier regressions seeking to explain U.S. inflation, the inclusion of *ROW* money does not significantly improve the explanatory power of the model.

IV. Conclusion

McKinnon has advanced an imaginative interpretation of macroeconomic events during the past decade of managed floating. In a number of important respects, however, his interpretation seems strikingly at odds with the empirical evidence. We conclude that a combination of sterilization techniques and exchange rate adjustments have provided central banks in the leading industrial countries with far more independence from monetary developments elsewhere than McKinnon's analysis suggests. Whether or not this judgment is correct, money supply developments in other leading industrial countries do not appear to have had a significant impact on U.S. inflation or on fluctuations in U.S. output since the advent of managed floating. Contrary to McKinnon, the evidence does not suggest that the Federal Reserve should alter its monetary stance in the light of monetary policies in the other leading industrial countries.

Acceptance of this proposition does not mean that the U.S. monetary authorities should ignore the behavior of the dollar in the exchange markets. An exceptionally strong *DM* led the Bundesbank to permit faster German monetary growth in 1978; for the same reasons, one can argue that an exceptionally strong dollar should have persuaded the Federal Reserve to expand U.S. money by more than it did during 1981–82. But such a judgement—whether valid or not—does not rest on the proposition that external money growth is an important determinant of U.S. inflation and economic activity under managed floating.

REFERENCES

Argy, Victor, *Exchange-Rate Management in Theory and Practice*, Princeton Studies in International Finance, No. 50, October 1982.

Bordo, Michael D. and Choudhri, Ehsan U., (1982a) "Currency Substitution and the Demand for Money," *Journal of Money, Credit and Banking*, February 1982, *14*, 48–57.

______ **and** ______, (1982b) "The Link Between Money and Prices in an Open Economy: The Canadian Evidence from 1971 to 1980," *Federal Reserve Bank of St. Louis Review*, August/September 1982, *64*, 13–23.

Carlson, Keith M., "The Lag from Money to Prices," *Federal Reserve Bank of St. Louis Review*, October 1980, *62*, 3–10.

Frenkel, Jacob A. and Levich, Richard M., "Covered Interest Arbitrage: Unexploited Profits," *Journal of Political Economy*, April 1975, *83*, 825–38.

Friedman, Milton and Schwartz, Anna J., "Money and Business Cycles," *Review of Economics and Statistics*, February 1963, *45*, 32–78.

Harberger, Arnold C., "A Primer on Inflation," *Journal of Money, Credit and Banking*, November 1978, *10*, 505–21.

Laney, Leroy O. and Willett, Thomas D. "The International Liquidity Explosion and Worldwide Inflation: The Evidence from Sterilization Coefficient Estimates," *Journal of International Money and Finance*, August 1982, *1*, 141–52.

Lehment, Herman, "Economic Policy Response to the Oil Price Shocks of 1974 and 1979: The German Experience," *European Economic Review*, May/June 1982, *18*, 235–42.

McKinnon, Ronald I., "Currency Substitution and Instability in the World Dollar Standard," *American Economic Review*, June 1982, *72*, 320–33.

Obstfeld, Maurice, "Can We Sterilize? Theory and Evidence," *American Economic Review Proceedings*, May 1982, *72*, 45–50.

Poole, William, "The Relationship of Monetary Decelerations to Business Cycle Peaks: Another Look at the Evidence," *Journal of*

Finance, June 1975, *30*, 697–712.

Shigehara, Kumihara, "Absorption of the Two Oil Shocks: The French Case," *European Economic Review*, May/June 1982, *18*, 249–61.

Solnik, Bruno, "An Empirical Investigation of the Determinants of National Interest Rate Differences," *Journal of International Money and Finance*, December 1982, *1*, 333–40.

Wonnacott, Paul, *U.S. Intervention in the Exchange Market for DM, 1977–80*. Princeton Studies in International Finance, No. 51, December 1982.

Deutsche Bundesbank, *Annual Report*, 1978.

International Monetary Fund, *International Financial Statistics*, various issues.

U.S. Department of Commerce, *Survey of Current Business*, various issues.

Faculty Ratings of Major Economics Departments by Citations

By PAUL DAVIS AND GUSTAV F. PAPANEK*

The predominant ranking criterion for judging the comparative quality of economics departments has been the faculty publication record in a set of "major" journals.[1] In this tradition is a recent article which generates rankings based on volume of pages published in twenty-four "top" journals during 1974–78 (Philip Graves et al., 1982). Separate rankings are calculated for total and per capita contributions, and credit is accorded to the institution at which the author is employed at the time of publication.[2]

I. Weaknesses of the Prevailing Ranking Criterion

This approach suffers from methodological and conceptual flaws:

Given faculty mobility, crediting articles by affiliation at the time of publication is bound to be inaccurate. The per capita ranking in the most recent article (Graves et al.) uses faculty affiliation in 1977. This penalizes departments which expanded rapidly in 1974–77, and overrates departments that lost productive faculty between 1974 and 1982.

Contribution to leading journals is an imperfect measure of research quality. It does not measure the critical and innovative aspects of articles. The contributions of those who use alternative forms of publication are ignored altogether.

The sociology of prestige determination in the economics profession can affect the ranking. Economists who focus on policy problems and/or interdisciplinary research have their professional achievements evaluated largely by individuals outside the profession. Since this threatens the profession's demarcation boundaries and prestige system, applied economists tend to be at the bottom of its prestige ladder. Conversely, those whose work is most divorced from outside contacts and evaluations—theoreticians and mathematical economists—are on the highest rungs of the ladder (see Axel Leijonhufvud, 1973, p. 329; John Kenneth Galbraith, 1971). The refereeing process in the top journals may then function as a screening device in a sociological, as well as in a quality maintenance, sense. A policy-related or interdisciplinary article of the same quality as a theoretical article will have a smaller probability of being accepted. Moreover, the prestige accorded to journals will be affected by their perceived orientation, with theoretically and quantitatively inclined journals ranked ahead of applied journals. This argument is supported by an experiment (Robert Hawkins et al., 1973) with a pair of dummy journals among a list to be ranked by respondents: the ficticious theoretical journal ranked in the top third; its applied counterpart finished in the bottom third. The prestige system of the profession thus can operate to influence article acceptance in a particular journal, as well as the choice of the journals to be surveyed, which may cause a theoretically oriented department to be ranked above an applied department of the same objective caliber in terms of analytical and innovative skills.

II. Citations: An Alternative Ranking Criterion

An alternative approach ranks Ph.D.-granting departments by the number of citations attributed to their faculty.[3] Citations include references to journals, books, and monographs. If a cited author is listed as a faculty member by two or more universities, full credit is given to each of the schools, a method common to publication attribution studies. The average number of citations for 1978 and 1981 is used to avoid any bias as a result of selecting a single year. However, there is remarkably little change in rankings between these years.[4]

---

*Graduate Student and Professor of Economics, respectively, Boston University, 270 Bay State Road, Boston, MA 02215. We are grateful to Augustin Llona, Lida Kyn, Oldrich Kyn, Peter Doeringer, and a referee for their help.

[1] John Bell and John Seater (1978), Donald House and James Yeager (1978), James Miller and Robert Tollison (1975), William Moore (1973), Albert Niemi (1975), John Siegfried (1972), and Pan Yotopoulos (1961).

[2] Opinion surveys, the principal alternative ranking method, are generally acknowledged to be highly biased.

[3] Drawn from the *Social Sciences Citation Index*, a periodical which compiles references from 5,200 social science periodicals of forty-five countries.

This citation-based ranking meets several criticisms of journal publication-based rankings:

Faculty names are taken from the latest available catalogues. Credit is therefore given to the institution at which a cited author is currently employed. This eliminates the distortions from crediting departments for articles on the basis of outdated faculty lists.

Citations measure the critical and innovative quality of an economist's work, as reflected by the degree to which fellow economists and other social scientists have found that work relevant to their own research. The publication criterion, in contrast, gives equal weight to all articles regardless of their impact on the profession.

The citation method also provides a better index of the contribution of published work, because it is less subject to the Leijonhufvud/Galbraith criticism. References are drawn from a wide variety of sources, not a small number of "prestigious" journals.

The citation-based methodology also suffers from defects. (*i*) The source credits only the first author in a multi-authored publication; others do not receive an accurate citation total. (*ii*) The contribution of economists who write survey articles may be exaggerated, if there is a tendency to cite such articles rather than the original works. (*iii*) Citations reflect work done over the course of a career and will overestimate the contribution of faculty whose creative contributions were made in the distant past. But the adjustment for "age" made below should correct for this. (*iv*) Inaccuracies arise in the citation-based method from padding or inertia, as retired or resigned faculty may continue to be listed for a year or two. But they are surely less than when credit is by affiliation at the time of publication. It is likely that any biases introduced by these weaknesses will average out for most departments. A more serious problem is (*v*) the absence of a quality control mechanism to guarantee a minimum standard, as the refereeing process does for articles in publication-attribution studies. Self-citations count as much as do the citations of others. The contributions of vainglorious economists will thus be overrated. More important, the citation-based rankings will overstate the contribution and quality of departments dominated by a controversial school of thought. They will garner many citations as their views are refuted. Often controversy will advance understanding, sometimes it will not. Excessive weight may thus be given to controversy in citation-based rankings, just as excessive weight may be given to limited theoretical contributions in rankings based on publication in leading journals. The two methods thus complement each other.

III. Ranking by Citation: The Results

Total citations (Table 1) is an index of total department impact, faculty depth and versatility. However, there are clearly great differences in department size. Faculties in both agricultural economics and economics were covered. State universities with two large departments therefore drop substantially in the rankings by citations per faculty member (also in Table 1), for example, Michigan State from 22 to 56 and Illinois from 21 to 38. The per faculty index is a better measure of average strength, while the total index measures depth of departmental strength, so both can be useful.

The total citation-based ranking of the top forty universities has been adjusted to reflect faculty "age" and citation dispersion. "Age," the number of years since a faculty member was awarded the highest degree, needs to be taken into account, or the criterion discriminates against departments with a young faculty of currently active scholars, who do

---

[4] Only five universities shifted by 5 to 7 places, one by more: the University of Colorado dropped by 11 places.

TABLE 1—UNIVERSITY RANKING, BY TOTAL AND PER CAPITA CITATIONS (1978 AND 1981)

University	Total Number of Citations[a]	Rank by Mean Number[b]	Mean Number
1. Chicago	5,778	1	156.16
2. Harvard	5,022	3	83.70
3. MIT	4,333	2	103.17
4. Stanford	3,998	6	59.67
5. Princeton	3,534	5	65.44
6. Yale	2,759	12	48.40
7. Pennsylvania	2,561	7	59.56
8. Wisconsin-Madison	2,216	14	54.22
9. Columbia	2,129	9	51.93
10. UC-Berkeley	2,099	15	44.66
11. Michigan	1,783	20	32.42
12. UC-Los Angeles	1,760	8	55.00
13. New York	1,584	11	49.50
14. Northwestern	1,511	13	47.22
15. Maryland	1,351	27	21.79
16. USC	1,294	17	34.97
17. Rochester	1,239	4	77.44
18. Minnesota	1,239	10	51.62
19. VPI	954	19	34.07
20. Washington	848	22	28.27
21. Illinois-Urbana	789	38	15.17
22. Michigan State	754	56	9.92
23. Duke	727	16	44.66
24. Massachusetts-Amherst	720	23	27.69
25. George Washington	717	39	14.63
26. Cal Tech	666	21	28.96
27. Cornell	623	24	23.96
28. Johns Hopkins	617	18	34.28
29. Claremont Graduate School	591	29	20.38
30. Virginia	577	30	19.87
31. Boston	569	43	12.93
32. Texas-Austin	507	40	14.08
33. UC-San Diego	502	53	10.73
34. Purdue	481	74	6.09
35. UC-Santa Barbara	471	32	18.11
36. Colorado-Boulder	466	37	15.53
37. Arizona	445	33	17.11
38. Washington-St. Louis	442	34	17.00
39. Brown	440	26	22.00
40. Ohio State	434	46	12.06
41. Vanderbilt	401	42	12.93
42. New School	399	28	21.00
43. CUNY	384	86	11.29
44. Texas A&M	376	31	18.80
45. Pittsburgh	371	60	9.27
46. North Carolina-Chapel Hill	364	58	9.58
47. Carnegie-Mellon	348	35	15.82
48. Iowa State	332	73	6.77
49. SUNY-Binghamton	328	48	11.71
50. Florida	314	54	10.47
51. Indiana	313	45	12.52
52. Rutgers-New Brunswick	298	59	9.31
53. New Mexico-Albuquerque	290	50	11.60
54. NC State-Raleigh	285	82	4.60
55. Arizona State	283	57	9.76
56. Houston	263	41	13.84
57. UC-Davis	252	47	12.00
58. Connecticut	248	65	8.27
59. SUNY-Buffalo	220	36	15.71
60. Pennsylvania State	212	69	7.85
61. Florida State	208	62	8.67
62. Southern Methodist	207	52	10.89
63. Wayne State	205	66	8.20
64. Rice	204	44	12.75
65. Georgia	196	67	8.17
66. Colorado State	184	94	3.07
67. Syracuse	182	61	9.10
68. Oklahoma State	176	92	3.38
69. Boston College	175	68	7.95
70. Oregon	173	63	8.65
71. Missouri-Columbia	172	93	3.18
72. Wisconsin-Milwaukee	167	55	10.44
73. Temple	153	90	3.73
74. SUNY-Stony Brook	140	71	7.37
75. Georgetown	140	80	4.67
76. Washington State	129	78	5.16
77. South Carolina	123	87	3.84
78. Utah	120	79	5.00
79. Tennessee	120	81	4.61
80. UC-Riverside	118	25	22.82
81. Ohio	117	72	7.31
82. Nebraska-Lincoln	114	75	5.70
83. Wyoming	105	70	7.50
84. Iowa	101	89	3.74
85. Notre Dame	100	84	4.35
86. Kentucky	100	86	3.85
87. Rensselaer	100	104	2.50
88. Case Western Reserve	95	64	8.64
89. American	95	76	5.59
90. Louisiana State	86	77	5.37
91. Illinois-Chicago Circle	83	83	4.37
92. Northern Illinois	83	88	3.77
93. Tulane	82	41	11.71
94. SUNY-Albany	82	91	3.42
95. Georgia State	82	95	3.03
96. Texas-Dallas	70	103	2.59
97. Montana State	69	97	2.87
98. Catholic	61	105	2.44
99. Cincinnati	52	106	2.36
100. Kansas	52	107	2.26
101. West Virginia	49	110	2.04
102. Tufts	48	96	3.00
103. Texas Tech	48	102	2.67
104. Saint Louis	47	85	4.27
105. Illinois State	46	98	2.87
106. Oklahoma	39	100	2.78
107. Rutgers-Newark	38	108	2.23
108. Utah State	35	115	1.25
109. So. Illinois-Carbondale	31	109	2.06
110. New Hampshire	31	112	1.72
111. Arkansas-Fayetteville	30	119	0.91
112. Northeastern	29	114	1.33
113. Howard	26	117	1.13
114. Mississippi	25	113	1.47
115. Kansas State	21	116	1.17
116. Clark	18	111	2.04
117. Miami	17	99	2.83
118. Lehigh	12	120	0.67
119. Bryn Mawr	11	101	2.75
120. Rhode Island	9	119	0.75
121. Fordham	8	121	0.44
122. Alabama-Birmingham	3	122	0.33

[a] When total departmental citations are equal, ranking is based upon citations per faculty member.
[b] When mean departmental citations are equal, ranking is based upon total departmental citations.

TABLE 2—UNIVERSITY RANKING CONTROLLING FOR AGE AND CITATION DISPERSION

	Number of Citations controlling for Age and Dispersion[a]	Rank Controlling only for Age	Rank Controlling for Dispersion
1. Harvard	500	1	1
2. Princeton	366	2	5
3. Chicago	364	3	2
4. MIT	277	5	3
5. Pennsylvania	271	7	4
6. Stanford	245	4	6
7. Yale	230	6	8
8. Wisconsin-Madison	214	9	7
9. UC-Los Angeles	201	11	10
10. Columbia	180	8	12
11. Berkeley	175	13	9
12. Michigan	168	16	11
13. Northwestern	160	16	11
14. Rochester	131	18	14
15. Washington	110	21	16
16. Maryland	107	12	19
17. Minnesota	103	20	13
18. Virginia	102	30	20
19. New York University	92	15	18
20. Illinois	88	24	21
21. USC	76	14	22
22. VPI	74	17	27
23. Michigan State	68	23	23
24. Cal Tech	68	25	26
25. Duke	66	36	17
26. George Washington	64	26	25
27. Johns Hopkins	60	29	29
28. Washington	60	39	24
29. UC-San Diego	59	27	33
30. Purdue	58	22	38
31. Boston	55	28	32
32. Massachusetts-Amherst	51	19	34
33. Brown	50	37	34
34. Cornell	48	32	28
35. UC-Santa Barbara	44	35	31
36. Arizona	42	33	36
37. Ohio State	37	38	35
38. Texas-Austin	36	34	37
39. Claremont	24	31	39
40. Colorado-Boulder	10	40	40

[a] Where the number of citations is identical, ranking was by magnitudes after the decimal point.

not yet have an extensive corpus of publications to be cited. Conversely, without adjustment, departments with an older faculty whose cited work may have been published decades ago, will be given an unduly favorable ranking. It is also desirable to adjust for dispersion because high dispersion can reflect a lack of intellectual depth. A department which has a single idea-generating giant is a less stimulating environment for students and for colleagues. This is likely to be reflected in future scholarship. Dispersion is controlled for by dividing the department total by the coefficient of variation; age by dividing it by mean number of years since the highest degree. Rankings adjusted for age, dispersion, and both are in Table 2.

The composition of the top 20 departments by total citations is quite close to that by publication-based rankings. However, below the top 20 there were many changes: 26 of the top 60 in Table 1 shifted 10 or more positions when compared to the total pages ranking in the Graves et al. study. Of 118 schools ranked by both methods, more than half (70) changed by 10 or more places. Some of the changes are far from trivial: Claremont and the New School improve by over 100 positions, Penn State declines by 35. Comparatively strong performances in the citation-based ranking reflect departmental strength in (*i*) publishing in books and monographs rather than journals, (*ii*) producing innovative and socially relevant research which is widely cited, and (*iii*) adding faculty in the recent past, not reflected in studies based on earlier publications. For instance, the State University of New York-Buffalo is theoretically oriented and falls in rank when the social science periodicals surveyed are broadened beyond the theoretical journals emphasized in prior studies. The improved ranking of the University of Massachusetts and Boston University reflects massive faculty additions in the recent past.

The combined control for age and citation variance generates interesting changes within the top 40 departments. The improvement in the positions of Brown, Virginia, Washington, and the University of Washington are traceable to the relatively low citation variance among faculty members. Claremont, Massachusetts, and University of Southern California decline in rank as a result of the skewness in their citation distribution: a single faculty member is responsible for more than 50 percent of all citations. The position of Texas-Austin and New York University drop as the result of both dispersion and age effects. The combined ranking of Table 2 is the best single indicator of relative faculty quality among the top 40 departments in our view.

The disparity between this citation-based ranking and those generated by previous studies illustrates the danger of excessive reliance on rankings based upon affiliation at time of publication, since these do not adjust for rapid changes in faculty composition of departments. It also illustrates the substantial impact which a moderate shift in the conceptual orientation of the ranking methodology can have on the final results. Despite its limitations, the advantages of the citation-based ranking methodology make it a viable alternative to the study of publications for determining faculty contributions to the development of ideas in economics. It provides an alternative means for assessing faculty quality.

REFERENCES

Bell, John G. and Seater, John J., "Publishing Performance: Departmental and Individual," *Economic Inquiry*, October 1978, *16*, 599–615.

Galbraith, John Kenneth, "The Language of Economics," in Andrea D. Williams, ed., *Economics, Peace and Laughter*, Boston: Houghton Mifflin Company, 1971, 32–44.

Graves, Philip E., Marchand, James R. and Thompson, Randall, "Economics Departmental Rankings: Research Incentives, Constraints, and Efficiency," *American Economic Review*, December 1982, *72*, 1131–41.

Hawkins, Robert G., Ritter, Lawrence S. and Walter, Ingo, "What Economists Think of Their Journals," *Journal of Political Economy*, July/August 1973, *81*, 1017–32.

House, Donald R. and Yeager, James H., "The Distribution of Publication Success Within and Among Top Economics Departments: A Disaggregative View of Recent Evidence," *Economic Inquiry*, October 1978, *16*, 593–98.

Leijonhufvud, Axel, "Life Among the Econ," *Western Economic Journal*, September 1973, *11*, 327–37.

Miller, James C. III and Tollison, Robert D., "Rates of Publications per Faculty Member in Forty-five 'Rated' Economics Departments," *Economic Inquiry*, March 1975, *13*, 122–23.

Moore, William J., "The Relative Quality of Graduate Programs in Economics, 1958–73: Who Published and Who Perished," *Western Economic Journal*, March 1973, *11*, 1–23.

Niemi, Albert W. Jr., "Journal Publication Performance During 1970–74: The Relative Output of Southern Economics Departments," *Southern Economic Journal*, July 1975, *42*, 97–106.

Dependency Rates and Savings: Another Look

By NATHANIEL H. LEFF*

In 1969, this *Review* published my paper which demonstrated that saving rates are inversely related to demographic dependency rates. Diverse studies by later researchers have confirmed this finding.[1] In his interesting paper (1982), Rati Ram has estimated my savings equations (as well as other equations) with a new data set: observations for a later period, and a set of countries that includes 54 additional less-developed countries (*LDC*s). Such replicatory work is important, and merits careful study. Many things happened in the world between 1964, the year of my observations, and 1970–77, Ram's sample period. Hence it would not be surprising if structural shifts also occurred in savings behavior. In fact, Ram's results are mixed. The equation he estimated utilizing my specification for a sample of 20 more-developed countries yields similar results: statistically significant, negative parameter estimates for the dependency terms (p. 540, Table 1, equation 18). For other samples and specifications the results differ; but as Ram states: "...using my data set, the closer one moves to the Leff sample and specification, the greater is the similarity of the estimates with the Leff results despite a substantial difference in the periods covered" (p. 539). Ram did obtain different results when he specified different equations and different samples. On the basis of those results, he suggests that higher dependency rates may not be associated with lower savings rates. Unfortunately, problems of model specification and sample heterogeneity make it difficult to interpret Ram's regressions.

I. Sample Heterogeneity

The sample in my 1969 paper included a total of 74 countries, of which 47 were *LDC*s. Subsequently, national income estimates for many more countries have become available. Ram's sample contains observations for 121 countries, of which he classified 101 as other than "industrialized." Serious questions arise as to whether these observations are sufficiently homogeneous to justify inclusion in the same sample for statistical inference. These questions apply both to Ram's overall sample of 121 countries, and to the various subsamples for which he has estimated regressions.

Many possible criteria may be used for stratifying country observations in studies of international savings behavior. The level of per capita income (which is highly correlated with other indices of general economic development) is one obvious possibility. One might also group countries in accordance with the dominant constraint on their economic growth. Another possible criterion for grouping is previous experience with respect to key economic variables like inflation, growth, and instability. Such country-specific experience helps determine expectations and institutions; and may thus also influence aggregate savings behavior. In addition, as Hollis Chenery and Moises Syrquin (1975) have demonstrated, countries' resource endowments and development strategies are systematically related to patterns of factor accumulation.

Another possible stratifying criterion is one that is often neglected in the literature, country size. That variable, too, influences development strategies, opportunity sets, and economic behavior.[2] Unfortunately, "size" is hard to model. Income levels, population, and territorial expanse have all been pro-

---

*Graduate School of Business, Columbia University, 403 Uris Hall, New York, NY 10027. I am grateful to Kazuo Sato for helpful comments and to Henry Woo for research assistance.

[1] See the studies discussed in my 1980 paper, p. 206. The idea that higher dependency ratios lead to lower saving was initially proposed in the context of household saving. For a discussion of why aggregate saving may be affected, see also pp. 206–08.

[2] On this point, see the theoretical and empirical analysis presented in B. Jalan (1982).

posed as a pertinent aspect. Each is relevant only to some degree; but efforts to develop a composite size index by applying principal-component or discriminant-function analysis have not been successful (B. Jalan, p. 20). However, even if we consider only the population dimension of size, we observe considerable heterogeneity within Ram's sample. In 1977, the mean size of the countries in his sample was approximately 23.1 million inhabitants, with a standard deviation of 59.1 millions. Some 72 percent of the countries had a population of 15 million or fewer inhabitants, the Chenery-Syrquin cut-off level for small size. And to accentuate the heterogeneity, fully 51 of Ram's countries, 42 percent of his sample, had a population of 5 million or fewer people.

The implications of this discussion are straightforward. As the number of countries with national income estimates has increased, researchers have been tempted to increase sample size by including all available observations (see Ram, p. 538). Statistical theory, however, makes it clear that more is not always better. Because of increasing heterogeneity, larger samples from disparate statistical populations may yield less information. Moreover, stratifying the sample on one dimension (for example, by income level, or on a developed country-*LDC* dichotomy) does not necessarily provide requisite homogeneity on other key dimensions.[3] Further, stratifying the countries on all important dimensions is no solution, for it leads to a drastic reduction in sample size. For this reason, international cross-section regressions may no longer be a fruitful research tool for approaching many questions in development economics.[4] By the same token, intrasample heterogeneity may prevent Ram's regressions from telling us much about the relation between dependency rates and saving rates. It is noteworthy that in the one Ram subsample for which greater homogeneity prevailed with respect to history, factor endowments, and economic policies (the industralized countries), the negative impact of the demographic variables on saving rates is clear (p. 540).

II. Specification Problems

The ways in which Ram specified many of his regression equations also pose some problems. Researchers working with international cross sections often specify the level of per capita income y, in an effort to isolate the impact of general development levels on the process they are studying. Ram may have carried this practice too far. In addition to y, four of his key equations include y^2 and/or an *LDC*-developed country dummy variable. This specification is regrettable, for it applies to all of the equations for Ram's full sample of less-developed and more-developed countries (p. 540). The main purpose of my 1969 paper was to show the relevance of the dependency variables for explaining the fact that saving rates are generally lower in *LDC*s than in more-developed countries. Ram's specification in the equations for his full sample effectively precludes use of his results from bearing on that central issue.

Interpretation of Ram's results is also difficult for other reasons. In half of his regression equations, Ram specified foreign capital inflow F as a determinant of saving rates. The theoretical basis for including F in a savings function is unclear, and is not explained.[5] Because F affects rates of return to capital, it may well belong in an *investment* equation; but including F in a savings function involves misspecification. Note further that the statistically significant parameter estimates Ram obtained for the F term tell us little about correct specification. For all we know, the equations that include F are investment functions rather than saving functions. And because of the high covariance between S and I in most countries, a high correlation between S and F can occur simply as a statistical artifact. Further, as economists have long noted, saving and investment are determined simultaneously. Consequently, single equation models of savings may lead to biased parameter estimates. Research on international saving behavior requires the same standards in specification and estimation as research on saving in the United States (myself and Kazuo Sato, 1975, pp. 1218–19). Ram recognizes this deficiency in his equations. As he expressed it: "A full scale treatment of the issue lies in the formulation and estimation of a simultaneous equations model which includes (besides the saving function) production, growth, birth rate, death rate, and possibly labor force participation equations. Such an exercise is beyond the scope of the present work" (p. 541). Kanhaya Gupta, however, has performed the necessary research. In 1975, he formulated such a simultaneous-equation model and estimated it using two-stage least squares. Gupta concluded that: "Leff's study underestimated the depressing effect of dependency rate on the saving rate" (p. 371).

---

[3] Note that the practice of stratifying by continent is of little help in this context. The theoretical basis of that criterion for stratification is unclear. Further, the countries located on individual *LDC* continents often show considerable heterogeneity with respect to the other variables mentioned earlier. Note further that a procedure of weighting country observations (for example, by income levels) would not deal with the problem of heterogeneity on other dimensions.

[4] The reader may wonder about the heterogeneity of the sample in my 1969 paper. Relatively small sample size, and the exclusion of observation for many "micro states" and extremely backward countries increased homogeneity. Some statistical information on homogeneity within that sample is reported in my 1971 paper, pp. 476; 479. In accordance with the considerations noted here, in my subsequent empirical research on savings I have worked with time-series equations for individual countries (and with simultaneous-equation rather than single equation models). See my articles with Kazuo Sato (1975; 1980).

[5] Ram notes (p. 538) that his specification of the savings function (including F) is "quite standard" in the development literature. But conventional practice does not justify misspecification, particularly where, as noted below, identification and simultaneity problems are also involved.

REFERENCES

Chenery, Hollis and Syrquin, Moises, *Patterns of Economic Development, 1950–1970*, Oxford, 1975.

Gupta, Kanhaya L., "Foreign Capital Inflows Dependency Burden, and Savings Rates in Developing Countries: A Simultaneous Equations Model," *Kyklos*, May 1975, *28*, 358–74.

Jalan, B., *Problems and Policies in Small Economies*, London: Croom Helm, 1982.

Leff, Nathaniel H., "Dependency Rates and Saving Rates," *American Economic Review*, September 1969, *59*, 886–96.

______, "Dependency Rates and Savings Rates: Reply," *American Economic Review*, June 1971, *61*, 476–80.

______, "Dependency Rates and Saving Rates: A New Look," in Julian L. Simon and Julie Da Vanzo, eds., *Research in Population Economics*, Vol. 2, Greenwich: JAI Press, 1980, 205–14.

______ **and Sato, Kazuo,** "A Simultaneous-Equations-Model of Savings in Developing Countries," *Journal of Political Economy*, December 1975, *83*, 1217–28.

______ **and** ______, "Macroeconomic Adjustment in Developing Countries," *Review of Economics and Statistics*, May 1980, *62*, 170–79.

Ram, Rati, "Dependency Rates and Aggregate Savings: A New International Cross-Section Study," *American Economic Review*, June 1982, *72*, 537–44.

Dependency Rates and Savings: Reply

By RATI RAM*

The two main concerns expressed by Nathaniel Leff in his comment will be discussed in the next two sections. As a preliminary, it may be useful to recall that the issue addressed in my 1982 paper is whether the empirical evidence supports the proposition that high birth (and death) rates, as reflected in a high proportion of the "young" population, depress aggregate saving rates in less-developed countries (*LDC*s). Therefore, questions concerning sample "homogeneity," specification of the saving equation, and estimation procedure should be considered with reference to their bearing on the principal issue investigated, and not in terms of general methodological propositions.

I. Sample Homogeneity

It is obvious that the consideration of sample homogeneity refers to the uniformity or constancy of the structural parameters across the observation set, and not to any other similarity in the characteristics of the sample countries. In fact, greater *heterogeneity* in the sense of a high variance in each regressor vector may be expected to yield sharper parameter estimates and might be an advantage. Although there are no fixed rules for investigating sample homogeneity of the kind that is relevant here (and an absolutely homogeneous sample may consist of just one observation), there are several precautions one can take to see that the principal results are not unduly sensitive to the composition of the sample: one can try to identify "extreme" observations and consider the effect on the estimates of their inclusion and exclusion; various subsamples based on some relevant characteristics can be considered to judge the sensitivity of the estimates; and sample observations may be varied with reference to some other criteria like the reference year.

Given the observed sensitivity of the results obtained in some earlier studies to the treatment of the "outliers" and the composition of the sample, great care was taken in my study to assess the sensitivity of the parameters of interest to sample variations. As stated (p. 538), the reported results are a fraction of those studied during the investigation. An attempt was made to identify unusual observations on the basis of both bivariate scatters and regression residuals, and the sensitivity of the estimates to the treatment of such observations was studied. Many different subsamples based on the level of development, geographical location, and inclusion in previous studies were examined. Several reference periods (besides the one reported) for different samples were explored. In no case was there any statistically significant adverse effect of the young dependency rate on aggregate saving in any group of *LDC*s and, in many cases, the relevant coefficient was observed to have a positive sign.

Therefore, there does not appear to be any reasonable basis for a concern about the interpretation of the results with reference to "heterogeneity" in the sample. In fact, the robustness of the estimates of interest with respect to variations in sample coverage, treatment of outliers, and the reference period, along with the comprehensive coverage of the sample, appear to provide greater assurance in regard to the reliability of the results than has been the case with previous studies.

Although population does not seem to be a relevant dimension for judging sample homogeneity, my sample is, on that dimension, not much more heterogeneous than those used in earlier studies. The sample used in the study by Leff (1969) seems as diverse; on the basis of population figures for 1977, 32 percent of the countries in the Leff sample had populations below 5 million and 62 percent had fewer than 15 million persons. In the 40-country sample studied by Kanhaya

---

*Department of Economics, Illinois State University, Normal, IL 61761.

Gupta (1975), 33 percent had populations below 5 million and 55 percent below 15 million. All are very "diverse" in this sense; but they reflect reasonably well the distribution of population size across countries in the world. It is not obvious that a more homogenous sample is even desirable; one would wish the sample to reflect the real world one seeks to study.

Indeed, population size probably does not matter for the issue of interest. Just to see how the estimate for the coefficient of the variable *DA* (proportion of young dependents) might vary with population size in the *LDC*s, I estimated linear and log-linear regressions similar to equations 4 and 5 of my Table 1 (p. 540) for four country groups: (a) all *LDC*s with per capita income below $2,000; (b) those with population below 5 million; (c) those with population below 15 million; and (d) those having more than 15 million persons. In every one of the eight cases, the coefficient is *positive*.

It is true that one *developed*-country sample shows the effect of young dependents to be significantly negative. However, despite all the homogeneity the sample might possess, it tells us nothing about the effect of the dependency rate in the *LDC*s. No manipulation of sample coverage in the *LDC* group so far seems to have produced a subsample that would show a significantly negative effect of dependency; even if there is one that does, it is not obvious what legitimate conclusion one can draw from that about the *LDC*s as a whole. Aside from everything else, there is an obvious tradeoff between sample homogeneity and the generalizability of the result to the less-developed world.

II. Specification and the Estimation Procedure

Three points may be considered here: inclusion of foreign capital inflows in some equations, insertion of the income-squared term or the *DC* dummy in some of the regressions, and the so-called simultaneous-equations bias.

Introduction of foreign capital inflows in aggregate saving equations is not new; nor is it entirely a matter of convention. A fairly extensive debate has been conducted in the savings literature about the process through which foreign capital inflows may reduce aggregate domestic saving and about the empirical evidence on the issue. Among others, J. L. Enos and Keith Griffin (1971), Griffin (1970), Griffin and Enos (1970), Earl Grinols and Jagdish Bhagwati (1976), Gupta (1970, 1975), Gustav Papanek (1972, 1973), Mohammad Rahman (1968), Bernard Wasow (1979), and Thomas Weisskopf (1972) have discussed the theoretical aspects or provided empirical evidence on the question. Given the rather extensive discussion of the issue in the literature, it does not seem fair to suggest that inclusion of the foreign capital variable makes the interpretation of my results difficult: exactly the opposite would appear to be a more reasonable position; that estimates obtained from specifications that completely ignore the variable are difficult to interpret especially in view of the likelihood of a positive covariance between the young dependency rate and foreign inflows and the consequent possibility of a downward bias in the estimated coefficient of the young dependency rate variable if foreign inflows are omitted.

It is also obvious from the results reported in Table 1 of my earlier paper and many other estimates that even if the foreign capital variable is omitted, one does not observe a significantly negative effect of the young dependency rate on savings in any *LDC* setting. Incidentally, the system of equations estimated by Gupta (1975) does have a foreign capital inflow variable in the saving equation, and the estimated coefficient for the variable has large negative values in both the structural and the reduced-form saving equations.

It is very difficult to see why inclusion of the income-squared term or the *DC* dummy in regressions based on both *LDC*s and *DC*s causes any problems or reflects a "regrettable" specification. Income variable in the specification is not just a stratification device to "isolate the impact of general development levels"; it is a genuine regressor believed to affect savings directly; it is not a proxy for something else. The income-squared term simply permits the effect of per capita income on the saving rate to be nonlinear, which seems entirely appropriate for samples in which the variance of income is large. The logarithmic specification clearly implies a nonlinear relation between income and saving. Thus, one should consider the income-squared term in linear specifications as not only unobjectionable but desirable; it should improve the estimates by avoiding a possible misspecification.

Also, there seems no reason for the view that the presence of the *DC* dummy somehow hinders an investigation of the effect of dependency on savings. The dummy is obviously a crude device for isolating the effect of any omitted variables that may differ across the two groups. It should sharpen the estimates. It is the specification without such a dummy that should be suspect because of a possible "omitted variables" bias.

As an empirical fact, even with the omission of the income-squared term and the *DC* dummy, along with the exclusion of foreign inflows, one sees no adverse effect of dependency on savings. Reestimating equations 2 and 3 of my earlier paper (Table 1, p. 540) after excluding the income-squared term and the *DC* dummy, as well as the foreign capital variable, one observes the following (with *t*-statistics shown in parentheses):

$$(1)\quad \begin{aligned} S = {} & \underset{(.50)}{7.01} + \underset{(.31)}{.085}\, DA - \underset{(-1.54)}{1.066}\, DB \\ & + \underset{(4.95)}{1.387}\, G + \underset{(7.91)}{.0039}\, PCI \\ & R^2 = .43,\ N = 121; \end{aligned}$$

$$(2)\quad \begin{aligned} \ln S = {} & \underset{(-.01)}{-.024} + \underset{(.04)}{.019} \ln DA - \underset{(-1.73)}{.376} \ln DB \\ & + \underset{(4.61)}{.242} \ln G + \underset{(7.96)}{.439} \ln PCI \\ & R^2 = .55,\ N = 110. \end{aligned}$$

Although these equations are probably misspecified due to the omission of relevant variables, the estimates provide an indication of the empirical insignificance of Leff's concerns pertaining to the inclusion of the foreign capital variable, income-squared term, and the *DC* dummy.

In my earlier paper, the question concerning the possible simultaneous-equations bias was handled in two ways. First, the test proposed by Halbert White (1980) for the joint hypothesis of homoscedasticity of the error term and no-specification error was applied to the basic *LDC* regressions, and the test statistics were found to be too small to justify rejection of the null. Second, using a reasonable set of instruments, of the kind implied by the system postulated by Gupta (1975), instrumental variable estimates (*IVE*) were obtained and reported in equation 20 (Table 1, p. 540). Like many of the ordinary least-squares estimates, *IVE* show a positive coefficient on the variable representing the young dependency rate. It is thus obvious that even though a single equation specification was employed, considerable care was taken to see that the main results are not affected by a possible bias in the estimates due to a "feedback" from the dependent variable or nonorthogonality of the regressors and the stochastic disturbance term.

III. Concluding Remarks

The objective of this reply has not been to argue that there are no problems in interpreting the estimates reported in my earlier paper. Some of the difficulties associated with such cross-section, country-level studies are obvious and well known. However, I do argue that, from the point of view of sample homogeneity, specifications used, or the estimation procedure employed, the estimates reported by me are no more difficult to interpret than the results claimed in earlier studies. On the contrary, considering the comprehensive coverage of the sample and the care taken to assess the sensitivity of the basic estimates to sample coverage, treatment of outliers, variations in the reference period, and the method of estimation, it would seem that the results reported in my earlier paper are more reliable than those of others, and it appears reasonable to believe that empirical evidence is lacking to support the view that higher dependency rates depress aggregate savings in the *LDC*s. Nevertheless, there is nothing infallible or final about these results. Further evidence may well warrant a different inference.

REFERENCES

Enos, J. L. and Griffin, Keith B., "Foreign Assistance: Objectives and Consequences: A Reply to Our Critics," *Economic Development and Cultural Change*, October 1971, *20*, 155–58.

Griffin, Keith, "Foreign Capital, Domestic Savings and Economic Development," *Oxford University Institute of Economics and Statistics Bulletin*, May 1970, *32*, 99–112.

______ **and Enos, J. L.**, "Foreign Assistance: Objectives and Consequences," *Economic Development and Cultural Change*, April 1970, *18*, 313–27.

Grinols, Earl and Bhagwati, Jagdish, "Foreign Capital, Savings and Dependence," *Review of Economics and Statistics*, November 1976, *58*, 416–24.

Gupta, Kanhaya L., "Foreign Capital and Domestic Savings: A Test of Haavelmo's Hypothesis with Cross-Country Data: A Comment," *Review of Economics and Statistics*, May 1970, *52*, 214–16.

______, "Foreign Capital Inflows, Dependency Burden, and Saving Rates in Developing Countries: A Simultaneous Equation Model," *Kyklos*, May 1975, *28*, 358–74.

Leff, Nathaniel H., "Dependency Rates and Savings Rates," *American Economic Review*, September 1969, *59*, 886–96.

______, "Dependency Rates and Savings Rates: Reply," *American Economic Review*, June 1971, *61*, 476–80.

______, "Dependency Rates and Savings: Another Look," *American Economic Review*, March 1984, *74*, 231–33.

Papanek, Gustav F., "The Effect of Aid and Other Resource Transfers on Savings and Growth in Less Developed Countries," *Economic Journal*, September 1972, *82*, 934–50.

______, "Aid, Foreign Private Investment, Savings, and Growth in Less Developed Countries," *Journal of Political Economy*, February 1973, *81*, 120–30.

Rahman, Mohammad Anisur, "Foreign Capital and Domestic Savings: A Test of Haavelmo's Hypothesis with Cross-Country Data," *Review of Economics and Statistics*, February 1968, *50*, 137–38.

Ram, Rati, "Dependency Rates and Aggregate Savings: A New International Cross-Section Study," *American Economic Review*, June 1982, *72*, 537–44.

Wasow, Bernard, "Saving and Dependence with Externally Financed Growth," *Review of Economics and Statistics*, February 1979, *61*, 150–54.

Weisskopf, Thomas E., "The Impact of Foreign Capital Inflows on Domestic Savings in Underdeveloped Countries," *Journal of International Economics*, February 1972, *2*, 25–38.

White, Halbert, "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, May 1980, *48*, 817–38.

Uncertain Innovation and the Persistence of Monopoly: Comment

By RICHARD J. GILBERT AND DAVID M. G. NEWBERY*

In her 1983 article, Jennifer Reinganum offers a counterexample to our earlier claim that preemptive patenting applies to situations where success in a patent race is uncertain. It is true that Reinganum's model includes uncertainty, and in her model an incumbent would not preempt competitors, but uncertainty is not the reason why her results differ from ours. Rather, the differences are the consequence of 1) a different assumption about timing of moves in the *R&D* game, 2) the absence of free entry, and 3) assumed diseconomies in the management of the firm. Although it is useful to call attention to the importance of these assumptions, it is equally important to understand why they lead to different conclusions.

In our 1982 article, we assumed that the incumbent had a first-mover advantage, and that he could credibly commit himself to a strategy that would make entry unprofitable. Potential entrants, observing this already chosen and committed strategy, would then choose to stay out, whether or not the subsequent race was deterministic or stochastic. In Reinganum's model, the entrant and the incumbent simultaneously choose their expenditure rates, and the incumbent has no first-mover advantage.

An alternative specification which has the properties of competition with a first mover is the bidding game described in our article (p. 517). In this game the firms bid for *R&D* services, and a firm can preempt by out-bidding its rivals. If competition among challengers establishes a certain bid, the incumbent can preempt by bidding slightly more than this amount. Preemption by the incumbent is a Nash equilibrium of this game, and the end result is similar to preemption by a firm with a first-mover advantage.

Preemption by the incumbent firm is a Nash equilibrium of the bidding game. Preemption is also a Nash equilibrium of the extensive form model when the incumbent can move first and commit to an *R&D* expenditure. The first-mover model follows the sequential perfect Nash equilibrium structure in Avinash Dixit (1980), although we had assumed free entry whereas Dixit models a duopoly. (The same equilibrium notion might be termed a Stackelberg equilibrium, which implies the ability of the leader—the incumbent—to commit to an *R&D* expenditure.) Reinganum's model is a Nash game where preemption need not occur, but the specification differs in extensive form and in the strategies available to competitors. These differences, and not the assumption of certainty or uncertainty, are primarily responsible for the contrasting conclusions in our papers. We return to the question of which is the more appropriate specification at the end of this comment.

I

Uncertainty is neither a necessary nor sufficient condition for the results demonstrated by Reinganum. Preemption can occur with an uncertain *R&D* process, and preemption need not occur when *R&D* is deterministic. For an example of the latter, consider a deterministic patent race as a one-shot Nash game. There is an incumbent and a challenger, and we employ the same notation as Reinganum: Π is the present value (*PV*) of the monopoly profits to the incumbent from the patent, π_I is the *PV* to the incumbent if the challenger wins, π_C is the *PV* to the challenger if he patents, *R* is the current profit of the incumbent, and *r* is the discount rate. The *R&D* technology is particularly simple. A firm can patent at date *T* and with a cost *X* (with $X \leq \pi_C$), or a firm can choose to do no *R&D*. The incumbent and the chal-

---

*Department of Economics, University of California, Berkeley, CA 94720, and Churchill College, Cambridge, England CB3 ODS, respectively.

lenger make their decisions simultaneously, as in Reinganum. Ties are resolved by flipping a fair coin. With this simple technology, pure-strategy Nash equilibria can exist where either or both firms engage in *R&D*. In addition, mixed-strategy equilibria can occur which have properties identified by Reinganum in her model of a stochastic *R&D* contest.

Let f be the probability that the incumbent patents, and g the probability that the challenger patents. Assuming (for simplicity of notation and without loss of generality) that T is sufficiently small for the time interval between the start of *R&D* and the patent date to be ignored, the challenger's expected return from *R&D* is

$$(1) \qquad V_C = (1-f)\pi_C + f\pi_C/2 - X.$$

The incumbent's expected return from no *R&D* is

$$(2) \qquad V_I^0 = (1-g)R/r + g\pi_I,$$

and the expected return from *R&D* is

$$(3) \qquad V_I^1 = (1-g)\Pi + g(\Pi+\pi_I)/2 - X.$$

In a mixed-strategy equilibrium, each competitor must be indifferent among the available alternatives. This requires $V_C = 0$ and $V_I^0 = V_I^1$, with the result that

$$(4) \qquad f = 2(1 - X/\pi_C);$$

$$(5) \qquad g = 2[\Pi - X - R/r]/[\Pi + \pi_I - 2R/r].$$

Both f and g are between 0 and 1 if $\pi_C > X > (\Pi - \pi_I)/2$. If the innovation is drastic so that $\pi_I = 0$, $\Pi = \pi_C$, then $g > f$ and the challenger patents with higher probability. By the same continuity arguments set out in Reinganum, this will continue to hold for a range of less drastic innovations.

Although the strategies in this game are stochastic, the *R&D* process is not. Thus uncertainty in the *R&D* process is not necessary for the results reported by Reinganum. (See Drew Fudenberg et al., 1983, for a more detailed analysis of competition with a deterministic *R&D* technology.)

In our article (Section IIIB, p. 520), we argued that provided the incumbent was risk neutral and that players held consistent expectations, our results continued to hold with stochastic innovation, of which Reinganum's formulation is a special case. Again, we stress that this is true given our assumptions about the specification of the game, free entry, and the absence of managerial diseconomies. Uncertainty is not sufficient to undermine the ability of the incumbent to preempt.

The second major difference between our earlier paper and Reinganum's is that we assumed free entry while Reinganum considers only the case of an incumbent faced by a single challenger. Free entry is important for several reasons. First, in the bidding model analog set out above, it ensures that the required size of the winning bid is well-defined. If the incumbent is to retain the market, it is not sufficient for him to probably deter a particular entrant, since there are many others, one of whom would be bound to enter. Instead he must outbid the most any entrant would be willing to pay. Free entry is not a necessary condition for preemption in the deterministic patent race of our earlier paper if the incumbent knows how much a challenger would pay to enter and if the incumbent can preempt by outbidding its opponent. In the deterministic patent race, the incumbent can preempt by spending no more than a rival's value, and hence no more than the cost to the incumbent from entry.

In general, preemption need not occur when competition is limited. It can be more profitable for an incumbent to accommodate a (not too rivalrous) competitor rather than deter entry, *if* deterrence requires dissipating profits to make entry unattractive. This result emerges, for example, in Dixit's model of limit pricing, and conditions for which entry prevention is desirable are discussed in Gilbert (forthcoming). Another example where entry deterrence may not be profitable is a stochastic patent race with a single opponent (as in Reinganum). The incumbent may be forced to spend more than the opponent to deter entry. Suppose a challenger would spend X, but the total *R&D* expenditure required to make the challenger's profits negative would be NX. An incumbent has to spend at least $(N-1)X$ to deter the single opponent and this can cost more than the profit lost from the entry of a single competitor.

Finally, for free entry to make sense in these models, there has to be some mechanism limiting the number of firms which can coexist. In Reinganum's model, if entry is profitable for any one challenger at an *R&D* expenditure level x_C, it is profitable for any additional challenger who enters at the same *R&D* intensity. A challenger's instantaneous expected profit at date t, conditional on no discovery before t, is (using her notation)

$$h(x_C)\pi_C - x_C, \tag{6}$$

and it is assumed that there exists an x_C such that this is positive. The present value of total expected profit for the challenger is

$$V^C(x_I, x_C) = \frac{h(x_C)\pi_C - x_C}{r + h(x_I) + h(x_C)}. \tag{7}$$

If expression (6) and hence the numerator in (7) are positive, there is no x_I that can make (7) negative. Additional expenditures on *R&D* by the incumbent decrease the probability that a challenger will successfully patent before any future date, but they do not lower the challenger's profit rate, conditional on no discovery. Since there are no fixed costs in the Reinganum model, a positive profit rate means that the challenger, and for that matter any other firm, can expect strictly positive profits from the patent race. With Reinganum's *R&D* technology, *entry deterrence is impossible.*

Given that an incumbent cannot keep out a potential competitor, it is not surprising that the incumbent acts in an accommodating way toward the challenger. Although it is useful to observe that an incumbent faced with unavoidable competition may invest less in *R&D*, it does not imply that the firm would not prevent entry if it could. Also, it is not reasonable to cite uncertainty as the cause of the accommodative response to entry, when the assumed *R&D* technology does not allow for entry deterrence at all.

The absence of managerial diseconomies is central to the effectiveness of a preemption strategy. Without a clear argument why diseconomies arise in the management of productive resources, it seems reasonable in our view to assume that any firm can perform the activities of any other firm or collection of firms in the same industry at equal cost. Building in managerial diseconomies, as Reinganum does, guarantees that preemption is a decreasing returns activity.

We can extend Reinganum's model in various ways to examine the effect of allowing entry. If we retain her original model but allow firms to choose the number of parallel *R&D* activities which they operate, then it is necessary to impose a limit on the number of such activities any one firm can manage, as well as a limit on the number of potential entrants to avoid an infinite number of activities competing for the infinitesimal period before invention. If so, then with Reinganum's Nash equilibrium concept, the incumbent firm and all rivals will operate the maximum number of activities, and if the incumbent is allowed to operate more activities than all of his rivals put together, then he will probably win. If the incumbent can increase the number of activities he can operate without bound, while the rivals remain limited, then he will preempt with probability one.

However, these results only go to show that the original model is quite unsatisfactory for examining preemption. If, instead, we modify the model so that all of the expenditure X is incurred at time zero, which then generates a success probability of h per unit of time for no subsequent expenditure, then there is a maximum number of viable programs $\bar{m}$, which have nonnegative expected profits. Again following Reinganum's notation, if the incumbent runs m parallel programs, and the n entrants each operate one program, the incumbent's expected profits are

$$V^I(m,n) = \frac{mh\Pi + nh\pi_I + R}{r + (m+n)h} - mX, \tag{8}$$

while each challenger has expected profits

(per program)

$$(9)\quad V^C(m,n) = h\pi_C/[r+(m+n)h] - X.$$

If the incumbent moves first and incurs the sunk cost mX, then it can credibly preempt, for it can choose an $\bar{m}$ such that $V^C(\bar{m},1) < 0$, which first occurs at the integer value

$$(10)\qquad \bar{m} = \mathrm{Int}\{\pi_C/X - r/h\}.$$

If the incumbent chooses $m < \bar{m}$, then the number of rival programs will be $n = \bar{m} - m$, and the incumbent's profits will be given by (8). The incumbent's monopoly profits Π can be written

$$(11)\qquad \Pi = \pi_I + \pi_C + L,$$

where L is the loss from competition. If $\bar{m}$ is large, then there is little inaccuracy in ignoring the integer constraint of (10), in which case (10) and (11) can be substituted in (8) to give

$$(12)\quad V^I(m,\bar{m}-m) = [mhL + \bar{m}h\pi_I + R]/[r+\bar{m}h].$$

This is increasing in m up to $\bar{m}$, and hence if the incumbent moves first he will choose $m = \bar{m}$ and deter entry, even though the innovation process is stochastic. Summarizing we have

PROPOSITION 1: *With free entry, no managerial diseconomies, and stochastic innovation, the incumbent's profits are maximized by committing to the entry-deterring R&D expenditure level whenever the loss from competition is positive.*

In this formulation the incumbent can commit by choosing to spend $\bar{m}X$ before the challenger can move, in which case no challenger would spend anything. Thus preemption is the Nash equilibrium of a game in which the incumbent firm can move first with a commitment.

If, on the other hand, the incumbent cannot move first, but must choose the number of programs m at the same time that the challengers choose their programs, then it is easy to see that preemptive *R&D* by the incumbent is not a Nash equilibrium. For at $m = \bar{m}$, $n = 0$.

$$\partial V^I(\bar{m},0)/\partial m = rh[\Pi - R/r]/(r+\bar{m}h)^2 - X.$$

Substituting expression (10) gives

$$\partial V^I(\bar{m},0)/\partial m = \{[r/(r+\bar{m}h)][(\Pi - R/r)/\pi_C] - 1\}X.$$

If the innovation is drastic, the incumbent's marginal profit is negative at $\bar{m}$, so that $m = \bar{m}$ cannot be a Nash equilibrium. Note that if the innovation is not drastic, preemption can be a Nash equilibrium in expenditure rates for the incumbent (without commitment) if the loss from competition is sufficiently large.

II

Preemption has to do with entry deterrence, and hence the conditions of entry and the timing of moves are of the essence. It makes little sense to consider preemption in a model in which assumptions are made that guarantee entry (i.e., by placing incumbent and challenger on equal footing and by specifying the number of firms and devising cost conditions to ensure interior equilibria as Reinganum does). If, on the other hand, the incumbent has the first move, and the number of entrants is endogenous to the model, then it becomes possible to deter potential competitors by reducing the benefits from entry to below the entry cost. If, in addition, there are no managerial diseconomies, then it is not only possible to deter entry, but desirable as well.

In any case, the assumption of a deterministic discovery function is neither necessary nor is it sufficient for the desirability of preemptive patenting. We have demonstrated that incentives for preemption persist when the discovery function is stochastic. We have also shown that with a deterministic discovery function and simultaneous Nash behavior, an incumbent cannot guarantee preemption. What Reinganum's model does illustrate is the importance of being able to move first and make credible commitments. Although preemptive patenting may be desirable, it need not be a *credible* response to potential competitors. The Nash formulation in the Reinganum model makes this clear. If preemption is successful, there will be no challengers. But if the challengers do no *R&D*, the incumbent is a monopolist in the *R&D* market and would not invest at the entry-deterring level.

Preemption is not a best response strategy in the Reinganum model, even though it might be desirable for the incumbent to preempt if it could. Whether preemption will occur ultimately rests on the extent to which firms can make prior entry-deterring commitments that are credible to potential competitors. We would argue that the incumbent has a natural temporal advantage since after all he is the incumbent, in which case the central issue is one of credibility. This in turn will depend on such questions as the extent to which *R&D* requires large irreversible investments (our paper, p. 517) and additional factors, not considered in this exchange, such as the scope of patent protection. We were careful in our original paper (Section IIIC) to point out that the incumbent might find it excessively costly to preempt against multiple threats, and we only argued that in some cases monopoly may persist. We see nothing in Reinganum's 1983 article to cause us to change this conclusion, although we feel that her contrasting results have helped clarify our understanding of the circumstances required for persistence.

REFERENCES

Dixit, Avinash, "The Role of Investment in Entry-Deterrence," *Economic Journal*, March 1980, *90*, 95–106.

Fudenberg, Drew, Gilbert, Richard, Stiglitz, Joseph and Tirole, Jean, "Preemption, Leapfrogging and Competition in Patent Races," *European Economic Review*, June 1983, *22*, 3–32.

Gilbert, Richard J., "Preemptive Competition," in F. Mathewson and J. Stiglitz, eds., *New Directions in Industrial Organization*, International Economic Association, forthcoming.

______ **and Newbery, David M. G.,** "Preemptive Patenting and the Persistence of Monopoly," *American Economic Review*, June 1982, *72*, 514–26.

Reinganum, Jennifer F., "Uncertain Innovation and the Persistence of Monopoly," *American Economic Review*, September 1983, *73*, 741–47.

Uncertain Innovation and the Persistence of Monopoly: Reply

By JENNIFER F. REINGANUM*

In my 1983 article in this *Review*, I reported results (based upon a stochastic model of invention) which were at variance with those (based upon a deterministic model of invention) reported previously in 1982 by Richard Gilbert and David Newbery (hereafter G-N).

In their comment, G-N claim that the differences in our respective results arise not due to the presence or absence of uncertainty, as I suggested, but due to (in my model) 1) a different assumption about the timing of moves in the *R&D* game, 2) the absence of free entry, and 3) assumed diseconomies in the management of the firm. The model in my paper was intentionally highly simplified, eschewing the issues of fixed costs and free entry, because the point was simple and intuitive. The point was *not* that preemption—or a weaker notion, stochastic preemption—would *never* occur. Nor did I claim that a deterministic invention process was either necessary or sufficient for preemption. My point was simply that when uncertainty is introduced into the G-N deterministic bidding model, the equilibrium outcome is quite different: in the deterministic model, the monopolist persists with probability one, while in the stochastic model, the monopolist will suffer entry—not just with positive probability—but with probability greater than one-half. Thus their conclusions regarding the likelihood of persistence are highly sensitive to the assumption of a deterministic invention process. Of course, this is not to say that their conclusions (or mine) are not equally sensitive to other modeling assumptions. In the sequel, I will examine the role of alternative assumptions regarding the order of play, free entry, and managerial diseconomies.

I. The Order of Play

There is no explicit discussion in G-N's 1982 article regarding the order of play in the game. However, indirect evidence tends to point toward simultaneous moves. For instance, "the strategy space for each firm is restricted to the research and development expenditure on product 2 and the price(s) the firm charges for the product(s) it sells" (p. 516). That is, monopolist and entrants alike each pick an investment level; in a sequential-move game, entrants would select a best-response function.[1]

Moreover, G-N assert that their model is formally equivalent to a particular type of auction model in which firms enter bids for the innovation; the winning bidder must then spend the amount of its bid on *R&D*. This is essentially a first-price auction with complete and perfect information. It turns out that in this framework, there is no need to specify timing conventions at all. It is easy to show that the simultaneous-move equilibrium is identical to the equilibrium specified by G-N.[2] Thus, regardless of the sequential or simultaneous nature of the bidding, the equilibrium outcome is the same. The agent with the highest valuation (in this case the incumbent monopolist) wins; however, the winner needs to pay only (a tiny bit more than) the amount of the next-highest valuation.

---

*Division of Humanities and Social Sciences, California Institute of Technology, Pasadena, CA 91125. The financial support of the National Science Foundation is gratefully acknowledged.

[1] In a related model, Partha Dasgupta and Joseph Stiglitz explicitly assume a leader-follower structure; they state that "active firms work on the reaction function of potential entrants; i.e., entertain von Stackelberg conjectures regarding their behavior" (1980, p. 10).

[2] Technically, if we think of this as a first-price auction, then both the sequential-move and simultaneous-move equilibria are really "*epsilon*-equilibria," since there is no *minimum* winning bid.

Thus it is perfectly reasonable to compare G-N's bidding model with a stochastic invention model in which firms move simultaneously. However, in evaluating their conjectures about the effects of uncertainty on their model, it is important to keep in mind their (implicit) sequential-move assumption. I will postpone until Section IV my comments on the appropriateness of the sequential-move framework.

II. Unrestricted Entry

Gilbert and Newbery emphasize that their model includes free entry,[3] that mine doesn't, and that this is (at least in part) responsible for our differing results. The issue of free entry is a red herring. In their original paper they claim that free entry is *not* crucial to their result (and, indeed, it isn't). Immediately following their main argument, they state that "...the same argument holds if competition for the patent is less intense, so that the potential entrant anticipates positive profits instead of the zero profits implied by [free entry]" (pp. 516–17). Moreover, in their model it is only the most efficient challenger which provides any competition for the incumbent; theirs may as well be a two-firm model.

Nevertheless, consider what happens in my model if one adds fixed costs and allows unrestricted entry (obviously both must be added at once; due to the assumption of decreasing returns to investment, with no fixed costs the number of firms is infinite). The fixed cost, denoted F, does not affect equilibrium flow investment, only whether or not the challenger firm plays. Since $F = 0$ implies infinitely many firms, and since the equilibrium challenger payoff with n firms can be shown to be decreasing in n (see my forthcoming paper), for any n one can find an F_n such that only the incumbent and n challengers want to play. For one challenger, this fixed cost is $F_1 = V^C(x_I^*, x_C^*)$. When n challengers play, the results are even stronger; now the incumbent firm invests less than *each* challenger. Thus its probability of persisting as a monopolist is less than $1/n$. Thus unrestricted entry can easily be accommodated without any weakening of my results.

III. Managerial Diseconomies

"Managerial diseconomies exist if the monopolist cannot conduct a research program or production plan as efficiently as any rival" (G-N, 1982, p. 518). This would not ordinarily seem to rule out decreasing returns to scale in the invention technology, as long as both incumbent and challengers alike are subject to the same decreasing returns. Actually, G-N want to say that if no managerial diseconomies exist, then the monopolist is as efficient as *all rivals put together*. Essentially, one needs to be able to run parallel *R&D* projects at no worse than constant returns to scale. Gilbert and Newbery "extend" my model to a number of parallel projects without including a fixed cost per project. Since this leads to an infinite number of parallel projects, they impose arbitrary upper bounds on the number of projects which may be undertaken by the incumbent and by challengers, and then discuss what happens when these bounds are differentially relaxed. In my earlier paper I reported results based on a model with parallel projects; the details of this model are available in an unpublished technical appendix. My own extension involved a fixed cost of K per project, and a flow cost, or research intensity, on each project. For simplicity, let the intensity be fixed at x, yielding hazard rate $h = h(x)$, and let n_I and n_C denote the number of parallel projects chosen by the incumbent and the challenger, respectively. The payoff functions now take the forms[4]

$$V^I(n_I, n_C)$$
$$= \left[n_I h\Pi(c) + n_C h\pi_I(c) + R - n_I x\right]$$
$$/\left[r + n_I h + n_C h\right] - n_I K$$

---

[3] I interpret the term "free entry" as meaning *unrestricted*, but not necessarily *costless*, entry.

[4] The present value of profits to a monopolist using the new technology with unit cost c are $\Pi(c)$, capitalized profits to the incumbent and entrant if the entrant

and $V^C(n_I, n_C) = [n_C h \pi_C(c) - n_C x] / [r + n_I h + n_C h] - n_C K,$

where it is assumed that $h\Pi(c) - x > 0$, so that the challenger has at least a chance at positive profits if the innovation is drastic. Differentiating V^i with respect to n_i, $i = I, C$ and simplifying yields the following necessary conditions at an interior Nash equilibrium (n_I^*, n_C^*).

$$(1) \quad [r + n_C^* h][h\Pi(c) - x] - h[n_C^* h \pi_I(c) + R] - KB^2 = 0$$

$$(2) \quad [r + n_I^* h][h\pi_C(c) - x] - KB^2 = 0,$$

where $B = r + n_I^* h + n_C^* h$.

PROPOSITION 1: *If the innovation is drastic and* $R > 0$, *then* $n_I^* < n_C^*$; *that is, the incumbent conducts fewer parallel projects than the challenger.*

PROOF:

If the innovation is drastic, then $\Pi(c) = \pi_C(c)$ and $\pi_I(c) = 0$. Combining equations (1) and (2) yields

$$(3) \quad (n_C^* - n_I^*)[h\Pi(c) - x] = R.$$

Since $h\Pi(c) - x > 0$, equation (3) requires that $n_I^* < n_C^*$.

Again, a simple continuity argument establishes that there is an open set of technologies which are not drastic, but for which Proposition 1 remains valid.[5]

IV. Conclusions

Gilbert and Newbery have argued that alternative assumptions regarding the conditions of entry, economies of scale and the order of play (and not uncertainty) are responsible for the differences in our respective results concerning the persistence of monopoly. It seems clear from the above discussion that at least the first two of these alternatives can be accommodated with no effect on the results. The models outlined in my original paper and in the preceding pages herein all describe circumstances in which, *were the invention process deterministic*, the incumbent would persist as the monopolist. But in the stochastic formulation, the incumbent enjoys a lower marginal benefit to invention than does the challenger when the innovation is drastic, or nearly so. Consequently, the incumbent invests less than the challenger and, *on average*, entry occurs.

All of my analysis is based upon interior Nash equilibria; that is, ones in which the challenger actually participates. Gilbert and Newbery object to assumptions which allow for the possibility of (or even guarantee) interior equilibria, saying that one cannot examine preemption and entry deterrence in such a framework. I think this takes an extremely narrow view of preemption and entry deterrence. When it is impossible to credibly preempt or deter entry with probability one, it still makes sense to ask whether stochastic preemption or stochastic entry deterrence are prevalent features of an industry. That is, are potential entrants discouraged *on average* from participating, or do they participate to a lesser extent than they would in the absence of the incumbent's strategic behavior?

I will concede that alternative assumptions regarding the order of play will typically yield different equilibrium results (G-N's deterministic bidding model is one example in which the order of play is of no consequence). But I do question the appropriateness of G-N's assumed order of play, and the attribution of the undesirable consequences to the patent system. In their comment, G-N extend their own model to the case of uncertainty, making explicit their sequential-move

---

patents the new technology are $\pi_I(c)$ and $\pi_C(c)$, respectively. Current flow revenues to the incumbent are denoted R. An innovation is *drastic* if it would drive the incumbent from the market; i.e., if $c \leq c_0$ where c_0 is the maximum level of unit cost such that $\pi_I(c) = 0$. In the interest of brevity, the reader is referred to my 1983 article for more complete definitions and the derivation of payoff functions.

[5] Assuming that success in *R&D* is a function of *fixed* rather than flow costs, Richard Freeman (1982) finds that a single domestic entrant will conduct more parallel projects than an incumbent foreign monopolist. Thus this result does not depend upon the fixed vs. flow specification of costs.

assumption. They show that in equilibrium the incumbent will choose an entry-deterring level of investment. However, this investment level will not be credible without some mechanism for commitment. They remark that

> Whether preemption will occur ultimately rests on the extent to which firms can make prior entry-deterring commitments that are credible to potential competitors. We would argue that the incumbent has a natural temporal advantage since after all he is the incumbent, in which case the central issue is one of credibility. [p. 242]

That monopoly power per se should confer a first-mover advantage seems debatable at best. One could argue equally persuasively the obverse claim that the potential entrant should have the first move, since the incumbent may not be aware of its existence or intent to invest until it actually does so. The entrant's investment alerts the incumbent to its presence, and it is the incumbent who must respond.

Even if one concedes a first-mover advantage to the monopolist, there remains the issue of credibility. In their original paper, G-N state that their purpose is to determine "whether institutions such as the patent system create opportunities for firms with monopoly power to maintain their monopoly power" (p. 514). However, if the difference in our respective results is due to their assumption that the incumbent can credibly commit itself to a preemptive investment level, then the responsibility for persistent monopoly clearly does not reside with the patent system, but with the (implicit) institution which facilitates an otherwise noncredible threat.

REFERENCES

Dasgupta, Partha and Stiglitz, Joseph, "Uncertainty, Industrial Structure, and the Speed of *R* and *D*," *Bell Journal*, Spring 1980, *11*, 1–28.

Freeman, Richard, "A Model of International Competition in Research and Development," manuscript, Federal Reserve Board, December 1982.

Gilbert, Richard J. and Newbery, David M. G., "Preemptive Patenting and the Persistence of Monopoly," *American Economic Review*, June 1982, *72*, 514–26.

______ **and** ______, "Uncertain Innovation and the Persistence of Monopoly: Comment," *American Economic Review*, March 1984, *74*, 238–42.

Lee, Tom K. and Wilde, Louis L., "Market Structure and Innovation: A Reformulation," *Quarterly Journal of Economics*, March 1980, *94*, 429–36.

Loury, Glenn C., "Market Structure and Innovation," *Quarterly Journal of Economics*, August 1979, *93*, 395–410.

Reinganum, Jennifer F., "Innovation and Industry Evolution," *Quarterly Journal of Economics*, forthcoming.

______, "Uncertain Innovation and the Persistence of Monopoly," *American Economic Review*, September 1983, *73*, 741–48.

Preemptive Patenting and the Persistence of Monopoly: Comment

By Stephen W. Salant*

In a recent article in this *Review* (1982), Richard Gilbert and David Newbery (G-N) advance the thesis that the patent system provides not only the opportunity but the *incentive* for a monpolist to preempt potential entrants to an industry. Indeed, they argue that—to maintain its monopoly—even an incumbent who is relatively inept in *R&D* would have an incentive to throw money at an *R&D* problem so as to beat more efficient potential entrants in the race to patent a new technology. At a Federal Trade Commission (FTC) conference held prior to publication of G-N's article, Gilbert emphasized the implications on their research for antitrust policy:

> This result is particularly onerous for economic efficiency. A firm could dissipate nearly all monopoly profits from preemptive patenting on wasteful *R&D* and still benefit from excluding competition. The productive activities are instead sacrificed in a wasteful *R&D* program whose main effect is to exclude the competition.... [1981, p. 236]

Gilbert and Newbery return to this theme in their article:

> Obviously if managerial diseconomies are significant, preemption is less likely to occur. What is more important is that in such cases the monopolist may dissipate much of the producers' surplus potentially available to the most efficient research group. *For it may still pay the monopolist to preempt more efficient rivals*, perhaps at the expense of almost all the potential profit.
> [1982, p. 518, emphasis added]

At the outset and conclusion of their article, G-N compare their analysis to the prior work of Oliver Williamson (1975), who reached similar conclusions about the persistence of monopoly. They contend that, in contrast to Williamson's analysis, their results do not depend on any assumptions about market imperfections.

In this comment, the characteristics which G-N attribute to the patent system are shown on the contrary to result either from faulty logic or from an implicit assumption that a market fails to operate. In particular, their main result is valid only under the assumption that, once a technology is patented, it is impossible for the patent holder to sell title to it or to sell a license on its exclusive use. If such transactions are permitted, entrants who can develop technologies more efficiently than the incumbent *always* win the patent race, preemptive patenting *never* occurs, and —while production will again be monopolized—there is no more reason to expect that the incumbent will acquire the entrant's patent than that the entrant will acquire the incumbent's patent.

I begin by reviewing G-N's argument, making explicit their assumption that no transfers of patent rights are possible. Next, it is shown that strikingly different conclusions follow if such transfers are permitted. Implications of the analysis for antitrust policy are considered in a concluding section.

I. The Incentives to Preempt and the Opportunity for Transfers of Patent Rights

The following notation will be used. Let π_d be the profits to each firm from duopolistic production following the patent race; π_m be the profit to any single firm from monopolistic production following the patent race; Δ be the *extra* cost an inefficient incumbent must expend to beat the most efficient potential entrant in the patent race; and $\Delta_{\max}$ be the excess of industry profits with monopoly over industry profits with duopoly ($\Delta_{\max} = \pi_m - 2\pi_d > 0$).

---

*Rand Corporation, 1700 Main Street, Santa Monica, CA 90406. Comments by Jonathan Cave, Glenn Gotz, Emmett Keeler, Charles Phelps, Herman Quirmbach, Rick Victor, and other members of Rand's Technological Innovation Workshop are greatly appreciated. I also thank Robert Clower, Richard Gilbert, David Newbery, and an anonymous referee for editorial assistance.

A. *Patent Rights Cannot be Transferred*

If the holder of a patented technology cannot sell title to it or license its exclusive use, then Gilbert and Newbery are correct in concluding that an inefficient incumbent has an incentive to preempt more efficient potential entrants. Such an incentive exists so long as the incumbent's profits from preemption exceed his profits from losing the patent race and then playing a duopoly game against the winner. The incumbent would earn π_d from the latter strategy. Since any entrant likewise[1] earns π_d if he wins the *R&D* race, in the competitive scramble against one another entrants will be willing to spend as much of π_d to win.[2] For the incumbent to preempt, therefore, he must spend Δ more than π_d, because of his assumed disadvantage in *R&D*. Such an expenditure of $\Delta + \pi_d$ would suffice to secure for him the subsequent monopoly profits, π_m. In short, the incumbent has an incentive to preempt if and only if $\pi_m - (\pi_d + \Delta) > \pi_d$. Hence G-N conclude that as long as an incumbent's inefficiency does not exceed $\Delta_{\max}$, he will choose to preempt.

B. *Patent Rights Can be Transferred*

In contrast, suppose 1) exclusive licensing is permitted, 2) adversaries are free to negotiate a mutually acceptable patent sale or licensing arrangement after the new patent is developed, and 3) both have complete information about the payoffs each will receive from producing as a monopolist or as a duopolist. In this case, an inefficient incumbent will *never* preemptively patent.

For in these circumstances, either the new entrant or the incumbent will acquire *both* patents on *some* (individually rational) terms and will become the monopolist of period two.[3] Total profits will be π_m. Without loss of generality, assume the net postrace payoff to the entrant is X if he wins the race and bargains. This could result either from his selling his newly patented technology for X or from his buying the incumbent's previously patented technology for $\pi_m - X$ and producing as a monopolist. Since any efficient entrant can anticipate receiving X if he wins the patent race, the competitive scramble will force the winner to be willing to spend X to secure the patent. If total industry profits are π_m and the entrant earns X if he wins the patent race and bargains, then the incumbent must earn $\pi_m - X$ if he loses the race to some entrant and bargains.

To preempt the entrant, the incumbent must spend $X + \Delta$. He would then earn $\pi_m - (X + \Delta)$ in profits. The incumbent preempts if and only if $\pi_m - (X + \Delta) > \pi_m - X$. Since this inequality never holds (for $\Delta > 0$), I conclude that an inefficient incumbent never preemptively patents no matter how the monopoly profits are divided in the second period. Furthermore, if the incumbent and the entrant are equally efficient ($\Delta = 0$), there is no more incentive for one firm to win the patent race than for the other. Finally, these propositions remain valid even if the incumbent and the entrant would earn different profits under duopoly.[4]

---

[1] To discover whether the patent system gives an incumbent any advantage over an entrant, assume that both entrant and incumbent would be in a symmetric situation if the entrant wins the race and both firms compete. If, instead, the incumbent and entrant would earn unequal profits if they compete, my claim that an inefficient incumbent would never preempt remains valid. For details, see fn. 4.

[2] Following G-N, I treat a patent race like an auction in which no costs are incurred by those submitting losing bids.

[3] It is inconceivable that duopoly would result if the parties bargaining had complete information and were unfettered by government regulation since that outcome can be Pareto dominated. If I departed from G-N's assumptions and instead assumed *incomplete information* about postrace payoffs, however, then it is conceivable that one could construct a model which supports their conclusion about preemption by an inefficient incumbent. This, of course, is a conjecture which can be evaluated formally.

[4] In the text, I have assumed that both the incumbent and the entrant would earn *equal* profits if a duopoly resulted. Alternatively, it might be assumed that the incumbent's profits as a duopolist ($\pi_{d'}$) would differ from the duopoly profits of the new entrant (π_d). This asymmetric case would occur, for example, if the technology under development was inferior to the incumbent's existing technology. In this more general case, the bargaining range is the interval $[\pi_d, \pi_m - \pi_{d'}]$. As long as $\pi_m - \pi_d - \pi_{d'} > 0$, the range is nonempty. For

The error in G-N's thinking is most clearly visible in Gilbert's (1981) presentation of their research. In a section specifically devoted to licensing, Gilbert discusses an example where $\pi_m = 150$ and $\pi_d = 70$. He appears to concede that preemption by an inefficient incumbent will not occur in the extreme case where the entrant garners *all* the bargaining surplus $(X = \pi_m - \pi_d)$ by making the final offer either to sell at $\pi_m - \pi_d$ or to buy at π_d; but he seems to contend that such preemption *will* occur as long as the entrant anticipates receiving anything less as the outcome of the bargaining:

> [Suppose] the patent is worth \$70 to a potential entrant and \$80 to the monopolist. A potential competitor could offer the monopolist exclusive patent rights and demand up to \$80 in compensation, since it is worth that much to the monopolist to avoid entry. The possibility of selling out to the monopolist increases rivals' incentives to invent, *but it does not eliminate the gains from preemption*. The patent is still worth \$80 to the monopolist and, depending on the outcome of the bargaining process, somewhere between \$70 and \$80 to a rival. If only for transaction costs, a rival should not expect to receive the full \$80 monopoly value of the patent.... [p. 222, emphasis added]

Assume transactions costs are zero but that the entrant cannot completely gouge the incumbent in the bargaining $(X < 80)$. Is there then any incentive for an inefficient incumbent to preempt by winning the patent race? It is true that the smaller the gains to the entrant from bargaining, the less he will strive for the patent and the more profits the incumbent will obtain by preempting. If, for example, the entrant earns a dollar less in the bargaining, he will be prepared to spend a dollar less in the *R&D* race and the incumbent will have to spend a dollar less to preempt. Hence the incumbent will earn a dollar more from preemption. But it is *also* true that the smaller the gains to the entrant if he wins the patent race and bargains the *larger* the gains to the incumbent from losing the race and bargaining. In my example, the incumbent would *also* earn a dollar more in profits if he *lost* the *R&D* race and then bought out the entrant at the assumed, more favorable terms. Since profits from preemption *and* profits from losing the race and buying out the entrant each increase by one dollar, the *relative* attractiveness of these competing alternatives does not change. Hence, if it is conceded that a more efficient rival will beat the incumbent in the patent race if he can garner *all* the surplus in the bargaining, it should likewise be conceded that he will beat the incumbent when he anticipates selling out at a lower price. This explains intuitively why I found algebraically that the advantage of preemption over losing and bargaining is *independent* of how π_m is divided in the bargaining.[5] Contrary to what Gilbert and G-N imply, the most trivial managerial diseconomy for the incumbent will *always* suffice to make losing the patent race and bargaining a better alternative than preemptive patenting.

If transactions costs are positive, preemptive patenting by an inefficient incumbent can occur—but *only* in cases where the inefficiency of the incumbent is *smaller* than the transactions costs. Let T denote the transactions costs dissipated in concluding the patent transfer. In many cases, these transaction costs amount to little more than the wages foregone if there is haggling, the cost of completing a standard form, and the cost of mailing it. Then $\pi_m - T$ remains to be divided. Let X again denote net postrace payoff to the entrant from bargaining. Then the incumbent must pay $X + \Delta$ to preempt. If, instead, he loses the race and bargains, he will receive $\pi_m - T - X$ (either spending $X + T$ to buy the entrant's patent or receiving $\pi_m - X - T$ from the sale of his own patent). The inefficient incumbent will preempt if and only if preemption is more profitable than

---

this asymmetric case as well, it is easy to prove that an inefficient incumbent will never preemptively patent. The proof of this proposition is identical to its proof in the text and need not be repeated. The reader need merely note that the proof in the text makes no mention of π_d or $\pi_{d'}$.

[5] How π_m is divided during the bargaining does have other effects. The larger the share anticipated by the potential entrants, the swifter the race and the larger the development costs of the victor.

bargaining with the entrant—that is, if and only if $\pi_m - (X + \Delta) > \pi_m - X - T, \Leftrightarrow \Delta < T$. It seems to me unlikely that *R&D* inefficiencies which are smaller than the transactions costs involved in a patent transfer will nonetheless be large enough to merit Gilbert's characterization of them as "onerous."

II. Policy Conclusions

With any stylized model, one hesitates to draw policy conclusions. Nonetheless, Gilbert did summarize this joint research to those in part responsible for U.S. antitrust policy at the FTC conference, and the paper was warmly received by Richard Craswell—the attorney with the Office of Policy Planning responsible for a written evaluation of it—as well as by other members of the Commission staff. Indeed, Craswell (1981, p. 279) was quick to *extend* what I regard as the fallacious preemption argument of Gilbert and Newbery to other contexts—for example, to a preemptive discovery of a new deposit by a natural resource monopolist. Under the circumstances, some comment on policy implications seems pardonable.

An inefficient incumbent has no incentive to preempt unless potential entrants anticipate that each will be prevented from bargaining with the incumbent if the entrant wins the patent race (or, following Craswell's analogy, if he wins the race to discover the resource deposit).

In fact, it may be reasonable for an entrant to anticipate that the antitrust authorities will prevent him from selling his patent to the incumbent or, alternatively, will prevent him from buying the incumbent's patent. Exclusive licensing is not per se illegal. But it has been regarded as "exclusionary" and contributing to a Sherman Act Section 2 violation. Anticipation of interference by the antitrust authorities would—in accordance with G-N's analysis—result in preemptive patenting and persistence of monopolies. But, as this comment has clarified, such preemptions should be recognized *as a consequence of antitrust interventions* and not as the inevitable concomitant of the patent system. Such antitrust interventions may have been *intended* to increase competition and lower prices to consumers by preventing the incumbent from acquiring an entrant's patent. But once such a policy comes to be *anticipated*, the more efficient inventors will *lose their incentive to win the R&D race* and the incumbent will retain his monopoly position by preemptive patenting (assuming $0 < \Delta < \Delta_{\max}$) instead of by purchasing a rival's license. In such cases, the antitrust policy fails to prevent monopolization; instead, it induces the inefficient incumbent to waste more of society's resources retaining his monopoly position and results in slower development of the new technology.

In cases where the inefficiency of the incumbent is sufficiently large ($\Delta > \Delta_{\max}$), the antitrust policy does increase competition—converting what otherwise would persist as a monopoly into a duopoly; but once again the policy retards the development of the new technology. To aggregate these consequences in order to evaluate properly the net impact of the antitrust policy on social welfare requires more explicit representation of the demand side of the model. Since no such representation is included in Gilbert and Newbery's article, it seems appropriate to conclude this comment here and to conduct the welfare analysis of the antitrust policy on another occasion.

REFERENCES

Craswell, Richard, "Comments on 'Patents, Sleeping Patents, and Entry Deterrence'," in S. Salop, ed., *Strategy, Predation and Antitrust Analysis*, Washington: Bureau of Economics, Bureau of Competition of the Federal Trade Commission, 1981, 271–86.

Gilbert, Richard J., "Patents, Sleeping Patents, and Entry Deterrence," in S. Salop, ed., *Strategy, Predation and Antitrust Analysis*, Washington: Bureau of Economics, Bureau of Competition of the Federal Trade Commission, 1981, 205–69.

______ **and Newbery, David M. G.,** "Preemptive Patenting and the Persistence of Monopoly," *American Economic Review*, June 1982, *72*, 514–26.

Williamson, Oliver, *Markets and Hierarchies: Analysis and Antitrust Implications*, New York: Free Press, 1975.

Preemptive Patenting and the Persistence of Monopoly: Reply

By RICHARD J. GILBERT AND DAVID M. G. NEWBERY*

Stephen Salant takes issue with a single paragraph of our 1982 paper (Section IIC, p. 518) and then claims that our conclusions result "either from faulty logic or from our implicit assumption that a market fails to operate." We respond by showing that our conclusions are logically consistent, and that Salant's argument is flawed by its incompleteness.

The preemptive patenting results hold whenever equally efficient firms compete with nonzero transaction costs (a point which is not contested by Salant), and whenever any cost disadvantage experienced by the incumbent is less than the transaction costs incurred in the process of bargaining for production rights. Furthermore, we will show that if the incumbent's cost disadvantage exceeds transaction costs, then the preemption result hinges on the *relative* transaction costs in the markets for *R&D* outputs and inputs. The conditions derived by Salant are erroneous, and indeed irrelevant to the incentives for preemption.

Salant mounts his main argument on the assumption that transaction costs are zero, but fails to realize the full implications of this counterfactual assumption. Ronald Coase told us many years ago that bargaining in the absence of transaction costs (broadly interpreted) will eliminate cost inefficiencies and yield a Pareto optimal allocation. By selecting excerpts from two different papers and then providing his own emphasis, Salant constructs the impression that our results contradict the Coase argument. We feel this creative journalism is an inaccurate representation of our conclusions.

The implications for preemptive patenting depend on transaction costs, as Salant shows. For any finite transaction cost, an equally efficient monopolist can preempt competitors (subject to the technological and strategic assumptions in our model). Salant argues that if a monopolist is less efficient than a rival, *and* if the cost penalty exceeds transaction costs, then the rival firm will preempt the monopolist. As it stands, this conclusion is false.

It is arbitrary to restrict bargaining, as Salant does, to a single market or group of markets. If firms can negotiate in the patent market, they can also negotiate in markets for inputs to the *R&D* process. Assuming (as we had explicitly done) no diseconomies that are solely due to management, technological inefficiency must be the result of inferior *R&D* inputs. These inputs can be bargained for as well.

I

Consider the case where a rival firm has, initially, a superior *R&D* technology. We shall examine the conditions that are necessary for the rival to succeed in obtaining from the incumbent an exclusive license on the existing product. Let X be the amount the rival would earn if it patents and bargains with the incumbent. Let Y be the amount the incumbent would receive by bargaining with the rival to transfer exclusive rights for the sale of the existing product. (We assume, as does Salant, that firms are equally efficient producers, so that only rights to sell, and not to manufacture, are important.)

The incumbent can remove the cost penalty by contracting with the suppliers of the superior technology. The incumbent can purchase the superior technology, either by contracting for the superior inputs or through purchase of the rival firm's assets. Let C be the cost for delivery of an innovation at date T. Since the cost net of transaction expenses should not exceed the innovation's value,

---

*Department of Economics, University of California, Berkeley, CA 94720, and Churchill College, Cambridge, England CB3 ODS, respectively.

$$C \leqq X + \Gamma_1, \tag{1}$$

where Γ_1 is the transaction cost for this contract. (The cost could be less with competition for the supply of *R&D* inputs. The minimum, assuming free exit of suppliers, is the cost of producing the innovation at date *T*, plus transaction costs.)

Let Γ_2 be the transaction costs incurred in a contract for the rights to produce both the new and the established product. Monopoly profits from the sale of both products, less transaction costs for the bargain, must satisfy the inequality

$$\Pi_m - \Gamma_2 \geqq X + Y. \tag{2}$$

The incumbent should sell an exclusive license to the rival firm, and not preemptively patent, if the incumbent would thereby receive more when the rival firm patents. This condition is

$$Y \geqq \Pi_m - C, \tag{3}$$

or, in view of (1),

$$Y \geqq \Pi_m - (X + \Gamma_1). \tag{4}$$

Substitute for *Y*, the incumbent's receipts with entry, in equation (2) and rearrange terms to find that the incumbent would benefit from the sale of exclusive rights only if

$$\Gamma_1 \geqq \Gamma_2. \tag{5}$$

The incumbent would sell out to a rival only if transaction costs in the market for *R&D* inputs, Γ_1, exceed transaction costs in the output market, Γ_2. If transaction costs in the market for *R&D* inputs are less than transaction costs in the output market (and less than the loss due to competition), the incumbent should bargain for the most efficient *R&D* technology and preemptively patent. If transaction costs exceed the loss due to competition, bargaining is ineffective. The incumbent would preemptively patent if its cost disadvantage is less than the loss due to competition, and would allow entry otherwise.

Thus the results in our paper hold with bargaining whenever transaction costs are large, or whenever transaction costs in the output market exceed transaction costs in the market for *R&D* inputs.

Once we admit bargaining, then if it occurs we find that the incumbent's incentive to preempt is *independent* of the incumbent's (*ex ante*) cost disadvantage. What matters are the relative transaction costs: the transaction cost incurred by the incumbent in the market for *R&D* inputs compared to the transaction cost incurred by a rival in the incumbent's product market. This follows because bargaining gives the incumbent an opportunity to overcome its initial cost disadvantage. Provided transaction costs do not exceed the loss from competition, the incumbent should preempt whenever the rival can expect to incur greater costs bargaining for exclusive rights to the incumbent's product than the incumbent would incur by bargaining for the rival firm's assets (prior to patenting).

It is not obvious to us that transaction costs always favor the rival, since it is quite common for innovative entrepreneurs to sell out to established firms.

II

We have shown that, contrary to Salant's assertions, (initial) cost disadvantages are not a determinant of the gains to preemption. What matters are relative bargaining costs. Of course, this discussion ignores other important determinants of the bargaining process, including uncertainty, asymmetric information, and the inability to write, verify, and enforce complete contracts. Also, we continue to stress that successful preemption requires many restrictive assumptions with regard to technological opportunities and the ability to commit to preemptive *R&D* strategies (see our comment in this issue on the paper by Jennifer Reinganum).

Even if the rival buys out the incumbent and preemptively patents, there is no change in the market structure, nor in the persistence of (a) monopoly. Salant's only point is that there may be cases in which *R&D* is done by a more rather than less efficient firm, thus avoiding some social waste, though he failed to specify the correct conditions under which this would occur.

Salant's conclusions are faulty because he only looked at bargaining for the incumbent's assets. If we admit that transaction costs are low enough to permit bargaining, then the bargaining may lead the incumbent to acquire the necessary *R&D* resources, preserving our results. The gains from preemption in a bargaining situation then rest on relative transaction costs.

REFERENCES

Coase, Ronald H., "The Problem of Social Cost," *Journal of Law and Economics*, October 1960, *3*, 1–44.

Gilbert, Richard J., and Newbery, David M. G., "Preemptive Patenting and the Persistence of Monopoly," *American Economic Review*, June 1982, *72*, 514–26.

______ **and** ______, "Uncertain Innovation and the Persistence of Monopoly: Comment," *American Economic Review*, March 1984, *74*, 238–42.

Reinganum, Jennifer F., "Uncertain Innovation and the Persistence of Monopoly," *American Economic Review*, September 1983, *73*, 741–48.

Salant, Stephen W., "Preemptive Patenting and the Persistence of Monopoly: Comment," *American Economic Review*, March 1984, *74*, 247–50.

Empirical Tests of the Life Cycle Hypothesis: Comment

By OWEN J. EVANS*

This comment examines the contention of Betsy Buttrill White (1976; 1978) that a multiperiod life cycle model *cannot* generate values of aggregate savings as large as in fact observed. I argue that her results reflect an entirely arbitrary choice of parameters and that for other equally plausible values of these parameters, the life cycle model generates saving similar to or greater than that actually observed. Hence, the model cannot be rejected, as White proposes, on the basis of such a simple test of plausibility. It should be noted at the outset that the argument of this paper is distinct from that of Lars Söderstrom (1982), who accepts White's basic result, and then shows that if the structure is generalized (to incorporate stochastic lifetimes), the result is not robust. My argument here is that White's contention is incorrect, when examined on its own terms.

I

White employs a constant proportional risk-aversion utility function to derive the individual desired consumption profile. She then makes assumptions about the present cross section of labor income, and future path of individual labor income to derive the level of saving for a representative individual in each age group. The actual population structure is then inserted to derive the implied total volume of personal saving. She justifies the assumption of a constant proportional risk-aversion utility function (see her fn. 2, p. 548) on the grounds that it is the *only* utility function consistent with the Modigliani-Brumberg homogeneity hypothesis, which is *wrong*. It is the only *separable* utility function consistent with it—a very different matter.

It is, of course, a tautology that in a zero-bequest life cycle model, the lifetime propensity to save is zero—for in the absence of inheritances and bequests, the individual ultimately consumes all his life income. *Aggregate* saving is positive in a life cycle growth model because of growth: population growth means that the young savers outnumber the old dissavers, and technological progress means that the young savers are relatively richer than the old dissavers. Population growth is present in the simulations conducted by White, but in each of her Tables 1–5, there is *no* technological progress—which makes them meaningless as a test of the life cycle hypothesis. Only in Table 6 is technological progress introduced, and here the life cycle hypothesis cannot be rejected on the basis of her results.

In her Tables 1 and 2, the desired consumption profile of the individual over life is flat, and the same is basically true in her Tables 3 and 4, apart from an adjustment to allow for the cycle of family size (through a consumption per person-equivalent mechanism). When the consumption profile is flat (or roughly so) but individual labor income rises with age, the younger cohorts dissave against future income growth leading to small or even negative aggregate saving, hence White's Tables 2, 3, and 4. In her Table 5 (still in the absence of technological progress), the desired consumption profile slopes upwards, and the simulated aggregate saving values are much more reasonable. White *dismisses* those cases with apparently plausible results because the parametric structure that generates plausible results for one year does not generate plausible results for another year (she is plugging in the actual population structure for each year), and because the implied rate of growth of individual consumption (7 percent per annum) is believed to be unreasonable for some unstated and

---

*Economist, Maritime Division, European Department, International Monetary Fund, Washington, D.C. 20431. Helpful comments were received from Barry Bosworth, Maxim Engers, Franco Modigliani, Irwin Friend, Robert Shiller, Patricia Stack, and Albert Ando. I alone am responsible for the views expressed or any errors.

TABLE 1—SAVING RATES AND WEALTH-INCOME RATIOS

Value of d	Saving Rate ($p = 0.00$)	Wealth-Income Ratio ($p = 0.00$)	Saving Rate ($p = 0.01$)	Wealth-Income Ratio ($p = 0.01$)	Saving Rate ($p = 0.03$)	Wealth-Income Ratio ($p = 0.03$)
0.5	0.36	10.68	0.30	8.89	0.11	3.09
1.0	0.22	6.44	0.17	4.87	0.04	1.18
1.5	0.15	4.33	0.11	3.13	0.02	0.50
2.0	0.11	3.14	0.07	2.18	0.01	0.16
3.0	0.06	1.87	0.04	1.20	–	–
4.0	0.04	1.21	0.02	0.69	–	–

Notes: Assumptions: $T = 55$, $T_1 = 45$, $g = 0.02$, $n = 0.015$, $r = 0.04$. Cases represented by – had negative aggregate saving and wealth. Saving rates and wealth-income ratios are measured relative to total income (including interest).

unexplained a priori reason. In her Table 6, in which at last technological progress is introduced, again several plausible values for aggregate saving are derived, and once again they are dismissed for somewhat confusing a priori reasons. She appears to be arguing that the parametric structure which (given the actual population structure for year A) works for year A should also work for year B (given its population structure). This attempt to generate realistic tracking is surely too stringent a requirement to impose on a simple perfect certainty model which lacks all kinds of real world complications. The model is steady state, with no allowance for uncertainty, for the business cycle or for changes in interest rates, so that realistic tracking *cannot* be expected. All one can do is to see whether the model can in principle account for aggregate saving within plausible empirical ranges. White does in fact find this, and her off-hand rejection of this favorable result is startling.

There are two facets of her structure worth pursuing further. She argues that one should assume an economic lifetime and earning span of 55 and 45 years (instead of 50 and 40, as in some other studies). She also proposes, following Malcolm Fisher (1956), that the need to provide for consumption of children should be incorporated through a consumption per person-equivalent mechanism—a modification that would also tend to lower simulated aggregate saving. The latter suggestion is perfectly reasonable, but I shall not pursue it here. Rather I shall present some simulation results in which an economic lifetime of 55 years and an earning span of 45 years are assumed. The broad range covered by the results is such that incorporation of the actual population structure (instead of assumed balanced population growth) and of consumption by children could not possibly change the flavor of the results.

II

The model used is identical to that of White, at the level of the individual. I assume that individual labor income grows at constant rate g per period until retirement after period T_1, after which no labor income is received.[1] Two further assumptions (that all individuals in the work force at a given period receive the same labor income, and that population growth proceeds at a constant rate n) permit the derivation of aggregate figures, such as the saving rate and the wealth-income ratio. Table 1 presents simulated aggregate saving rates and wealth-income ratios with economic lifetime (T) of 55 years and earning span (T_1) of 45 years.

In Table 1, aggregate saving rates are presented for a range of values of the two parameters of the utility function assumed by White (namely the pure time preference parameter, p, and the elasticity of marginal utility with respect to consumption, d). The table demonstrates the sensitivity of the results to the choice of parameters. Yet in her

---

[1] See ch. 2 of my 1982 dissertation for details. Notation is as in White (1978), unless otherwise stated.

TABLE 2—MEASURES OF AGGREGATE SAVING[a]

Year	White's Measure (1)	*NIPA* Personal Saving (2)	*NIPA* Disposable Income (3)
1953	61.4	20.3	276.4
1959	72.3	21.2	338.4
1964	92.7	29.0	432.4

[a] All figures are in billions of 1958 dollars. Columns (2) and (3) are from the *National Income and Product Accounts, 1929–76*, and consist of the current price figure, deflated using the personal consumption deflator, adjusted so that 1958 is the base year.

simulation, White seems to have used a single value for d, namely unity. The simulated saving rate is seen to range from negative to as high as 36 percent and correspondingly wealth-income ratios range from negative to as high as 10.7.

It might be argued that the real interest rate assumed (4 percent per annum) is too large—but plausible figures can still be obtained with smaller assumed rates of interest. For instance, with an interest rate of 2 percent per annum, zero time preference rate, a log utility function ($d = 1.0$), and the other assumptions as before, the aggregate saving rate is 13 percent and the wealth-income ratio is 3.94.

Plausible figures for the U.S. household saving rate are 5 to 15 percent, and for the wealth-income ratio, 3.5 to 4.5 depending on the definition (see the discussion in my dissertation). The range of results developed above is large enough that White's critique, and the resulting conclusion that the life cycle model cannot reasonably explain the level of aggregate saving in the United States, can simply be dismissed. This does not, of course, say anything as to whether the life cycle hypothesis is *empirically* relevant, but it does tell us that it cannot be discarded on the basis of such simple theoretical tests of plausibility.

III

A further reason[2] for White's argument being dubious must be mentioned. The measure of actual aggregate saving which she employs is severely flawed in that it includes in saving purchases of consumer durables, but does not subtract their depreciation.[3] Consequently, her estimate represents *gross saving*, not net saving, which is presumably the appropriate theoretical concept for testing the life cycle hypothesis.

Gross saving is, of course, a severely upward biased measure of the relevant concept, net saving, as is illustrated in Table 2. In each of the three years examined by White, her measure is roughly three and one-half times larger than conventionally defined *NIPA* saving. She had noted that, "At best, the simulated values are about 60 percent of the observed values" (1978, p. 547). Consequently, since the observed values employed are several times inflated, her results can in fact simply be dismissed, since the simulated values fail to reject the life cycle hypothesis when compared to a properly computed benchmark. This is even without taking account of the earlier criticisms of the construction of the simulated values. Certainly, any suggestion that the conclusions of the analysis stand when more appropriate saving data are used as benchmark would need to be more than pure assertion. It should also be noted that her attempt to require realistic tracking, aside from its dubious nature in terms of methodology, is meaningless as the benchmark data used for checking the tracking properties are not appropriate. Accordingly, the fact that her inflated measure of saving exceeds her simulated values can throw no useful light on the adequacy of the life cycle model.

---

[2] First pointed out to me by Franco Modigliani.

[3] The estimates are taken from the *Flow of Funds Accounts 1949–78*, rather than the conventional *NIPA* accounts. No justification is provided for the use of gross rather than net saving.

IV

The contention by White that the life cycle model is incapable of generating aggregate saving as large as observed in the United States is shown to be based on arbitrary parameter choice, and spurious choice of benchmark data. I have shown that the model is in fact capable of generating saving rates smaller than, equal to, or larger than, those observed empirically. In particular, the failure by White to incorporate technological progress in the bulk of her simulations makes these simulations meaningless. In those which incorporate technological progress, the life cycle model cannot reasonably be rejected. Finally, there is the incorrect use of "gross savings" as benchmark to which the simulated values are compared; if a correct benchmark had been employed, even the simulated values (with all their inadequacies, as discussed above) would not have rejected the life cycle model.

REFERENCES

Evans, Owen J., "The Life Cycle Inheritance: Theoretical and Empirical Essays on the Life Cycle Hypothesis of Saving," unpublished doctoral dissertation, University of Pennsylvania, May 1982.

______, "Tax Policy, the Interest Elasticity of Saving, and Capital Accumulation: Numerical Analysis of Theoretical Models," *American Economic Review*, June 1983, *73*, 398–410.

Fisher, Malcolm, "Explorations in Savings Behaviour," *Oxford Institute Economics Statistics Bulletin*, August 1956, *18*, 201–77.

Söderstrom, Lars, "The Life Cycle Hypothesis and Aggregate Household Saving," *American Economic Review*, June 1982, *72*, 590–96.

White, Betsy Buttrill, "On the Rationality of Observed Saving: A Critique of the Life Cycle Hypothesis," unpublished doctoral dissertation, Stanford University, 1976.

______, "Empirical Tests of the Life Cycle Hypothesis," *American Economic Review*, September 1978, *68*, 547–60.

Board of Governors of the Federal Reserve System, *Flow of Funds Accounts 1949–78*, Washington: USGPO, 1979.

U.S. Commerce Department, *National Income and Product Accounts 1929–76*, Washington: USGPO, 1981.

Empirical Tests of the Life Cycle Hypothesis: Reply

By BETSY BUTTRILL WHITE*

My 1978 paper presented simulations of the life cycle model as a test of the model's ability to explain aggregate personal saving in the United States. A number of simulations, reflecting a wide range of assumptions and incorporating actual demographic and income data, were presented for each of three "test" years. The major conclusion drawn from the analysis was that saving for future consumption does not account for the totality of observed aggregate personal saving.

The comment by Owen Evans contains two major criticisms of this analysis. The first is that the analysis uses inappropriate saving data to measure observed personal saving in the three test years. On this point, Evans is correct and it is important that this be pointed out. However, the major conclusion of the analysis stands when the simulation results are compared with more appropriate saving data.

Evans' second major criticism is that the results of the analysis "reflect an entirely arbitrary choice of parameters." The choice of parameter values (the rate of time preference, the elasticity of marginal utility with respect to consumption and the real rate of return) determines the *shape* of the optimal lifetime consumption paths employed in the simulations. However, for the utility function used, a single optimal consumption path represents an infinite number of *sets* of parameter choices.[1] Consequently, if the 1978 analysis is lacking, it is because it tested an insufficient number of "optimal" consumption paths, not that it relied on an arbitrary choice of parameters. The consumption paths employed in the simulations were illustrated in my Figure 1 (1978). Given the range of consumption paths spanned by the analysis it would seem that such a deficiency is unlikely.

Evans presents additional simulations using parameter values different from those employed in the 1978 analysis and does not incorporate actual income or population data. The results are qualitatively the same as those of my simulations in that they demonstrate that, in Evans' words, "the model is...capable of generating saving rates smaller, equal to, or larger than, those observed empirically" (p. 257). Such a claim leads me to believe that the true area of disagreement is the degree of stringency which one deems necessary or appropriate for testing theoretical models.

In reference to my earlier article, Evans states: "[White] appears to be arguing that the parametric structure which (given the actual population structure for year A) works for year A should also work for year B (given its population structure). This attempt to generate realistic tracking is surely too stringent a requirement to impose on a simple perfect certainty model which lacks all kinds of real world complications" (p. 255). I agree that to test the original life cycle hypothesis with all of its simplifying assumptions, to find it lacking, and then to claim it is inadequate, would be unreasonable. However, the simulations in the 1978 article and additional supporting simulations (see my 1976 dissertation) employ progressively more realistic assumptions and actual income and demographic data. The use of these more realistic assumptions (including growth in income, impact of families, and imperfect capital markets) and actual data does not improve the model's ability to explain saving across all three test years. Given these efforts to enhance the original "simple perfect certainty model," it does not seem unreasonable

---

*Manager, Banking Studies Department, Federal Reserve Bank of New York, Thirty-Three Liberty Street, New York, NY 10045. The views expressed herein are my own and not those of the Federal Reserve Bank of New York or the Federal Reserve System.

[1] Evans asserts that utility functions other than the one employed in the 1978 article are consistent with the life cycle hypothesis. Again, the utility function is key to the simulations only to the extent that it determines the optimal consumption path.

to expect the enhanced model to show consistency across years in explaining what it purports to explain.

REFERENCES

Evans, Owen J., "Empirical Tests of the Life Cycle Hypothesis: Comment" *American Economic Review*, March 1984, *74*, 254–57.

White, Betsy Buttrill, "On the Rationality of Observed Saving: A Critique of the Life Cycle Hypothesis," unpublished doctoral dissertation, Stanford University, 1976.

______, "Empirical Tests of the Life Cycle Hypothesis," *American Economic Review*, September 1978, *68*, 547–60.

ERRATA

Declining Industries and Political-Support Protectionist Motives

By Arye L. Hillman*

As noted in the text following equation (10), p. 1184, in my paper published in this *Review*, December 1982, the expansion $\tilde{M}_{pp^*}$ is positive, hence yielding the result that authorities guided by political-support protectionist motives respond to a decline in the world price of a domestic industry's output by permitting the industry's domestic price inclusive of the politically adjusted tariff to fall. As printed, however, the expression for $\tilde{M}_{pp^*}$ in equation (10) was incorrect, and appears as not unambiguously positive. The correct expression is

(10)

$$\tilde{M}_{pp^*} = \underset{(+)}{\Pi_p}\;\underset{(-)}{\Pi_{p^*}}\;\underset{(-)}{M_{\Pi\Pi}} + \underset{(+)}{\Pi_p}\;\underset{(+)}{V_{p^*}}\;\underset{(+)}{M_{\Pi V}} + \underset{(-)}{V_p}\;\underset{(+)}{V_{p^*}}\;\underset{(-)}{M_{VV}} + \underset{(-)}{V_p}\;\underset{(-)}{\Pi_{p^*}}\;\underset{(+)}{M_{\Pi V}} > 0,$$

which confirms the arguments in the paper.

---

*Department of Economics, Bar-Ilan University, 52 100 Ramat Gan, Israel.

NOTES

The eleventh annual AEA Summer Program for Minority Students will be held June 18–August 10, 1984, at the University of Wisconsin-Madison. The Program is open to black, Hispanic, and native American graduate students in economics, and offers intensive instruction in the core areas essential for graduate study. Students need not be economics majors, but must have completed one year of economic principles. Preference will be given to those who have also completed one year of calculus. Although the program is designed primarily for students completing the junior year, sophomores and seniors are also encouraged to apply. Participants will receive tuition, room and board, transportation, and a cash stipend. For further information and application materials, write to AEA Summer Program, Department of Economics, University of Wisconsin, 1180 Observatory Drive, Madison, WI 53706 (telephone 608+263-2441).

The Pennsylvania Conference of Economists will meet May 31 and June 1 in Allentown at Muhlenberg College. For further information, contact Kenneth G. Ainsworth, Vice President–Program, Allegheny College, Meadville, PA 16335.

The Canadian-American Business Studies Project will hold a forum, June 3–5, "Business/Management Issues in Canada," in East Lansing, Michigan. Topics are Business Regulations, Management of Natural Resource Industries, Foreign Market Development, and Transportation Management. Interested American scholars are urged to inquire as enrollment is limited (lodging and most food will be provided). Contact Professor Ronald Savitt, Department of Marketing and Transportation Administration, Graduate School of Business Administration, Michigan State University, East Lansing, MI 48824.

A Centre for Economic Policy Research has been established in London to promote independent analysis and public discussion of open economies and relations. The Centre will coordinate selected academic work already underway, develop new research initiatives, and channel resources. The Director is Professor Richard Portes. The Centre will not be an in-house research institute—Research Fellows and Associates will be invited to take part in particular programs. Publications will include a Discussion Paper series and periodic *Bulletins*. For further information, contact Mrs. Wendy Thompson, Administrative Director (Programme and Finance), Centre for Economic Policy Research, 6 Duke of York Street, London SW1Y 6LA.

The Bureau of the Census has published *Census '80: Product Primers* (1–7) to help college instructors, students, and researchers use the 1980 census products. To receive review copies of each *Primer*, telephone Les Solomon, College Curriculum Support Project (301+763-2370), or write CCSP Mailing List, User Training Branch, Data User Services Division, Bureau of the Census, Washington, D.C. 20233.

The Kansas Collection of the University of Kansas Libraries announces that the Jabez Bunting Watkins Collection is open to research. It is comprised of correspondence and business records which chronicle the diverse entrepreneurial interests of J. B. Watkins, and is of national significance because it details many phases in the growth and development of the trans-Mississippi midwestern and south central United States (1871–1946). Address inquiries to Kansas Collection, University of Kansas Libraries, Lawrence, KS 66045.

A Society for the Advancement of Behavioral Economics (SABE) was recently established to advance the course of interdisciplinary research in economics, and to provide an opportunity for scholars and practitioners to exchange views and report on research. The SABE will hold its first conference in the United States in May 1984. Further information may be obtained from the executive director, Dr. Benjamin Gilad, Department of Economics, Rutgers University, Newark, NJ 07102, or Professor Shlomo Maital, Department of Economics, Technion, Haifa, Israel.

The Trustees of the Rees Jeffreys Road Fund, through the *Journal of Transport Economics and Policy*, are offering a prize of £1,000 for an essay in "Transport and Information Technology." The winning essay will be published in the *Journal*. Submissions must neither have been previously published nor submitted for publication elsewhere, and must not exceed 15,000 words, including appendices if any. Each entry must be accompanied by an official entry form which is available with further information from the Journal Secretary, University of Bath, Claverton Down, Bath BA2 7AY (telephone: 0225-61244). The deadline is September 28, 1984.

The Social Science Research Council has established a Committee on States and Social Structures. The Committee is completing an edited volume of essays, *Bringing the State Back In*, to be published in 1984. For further information, contact Martha A. Gephart, SSRC, 605 Third Avenue, New York, NY 10158, and ask to be placed on the mailing list.

Call for Papers: The second Annual Meeting of the Association of Managerial Economists will be held in Dallas, Texas, December 28–30, 1984. Three sessions of contributed papers will be featured in conjunction with the ASSA meetings. Areas of interest include Forecasting, Pricing, Advertising, Diversification, Competitive Strategy, Innovation, Financial and Location Decisions, Regulation and Antitrust Policy. Both members and nonmembers are invited to submit papers and/or make program suggestions to the Program Chair, Professor Mark Hirschey, Graduate School of Business, University of Wisconsin, Madison, WI 53706.

The annual Conference of the Union for Radical Political Economics (URPE) will be held in Dallas, Texas, December 28–30 as part of the ASSA meetings. The URPE Program Committee invites proposals for individual papers and sessions. Please send abstracts (two copies) no later than April 30 to Fikret Ceyhun, Department of Economics, University of North Dakota, Grand Forks, ND, 58202.

Call for Papers: The Illinois Economic Association will hold its fourteenth annual meeting, October 26–27, 1984. Papers should be sent by June 1 to Professor Bernard McCarney, Department of Economics, Illinois State University, Normal, Illinois 61761 (telephone 309 + 438-2106).

Call for Papers: The European Finance Association will hold its eleventh annual meeting, August 30–September 1, 1984, at the Manchester Business School. To serve as a chairperson, discussant, or to present a paper (copies or detailed abstracts must be sent by April 1), contact Professor Richard C. Stapleton, Manchester Business School, University of Manchester, Booth Street West, Manchester M15 6PB England.

Call for Papers: The Association for Cultural Economics will hold the third International Conference on Cultural Economics and Planning, April 25–27, 1984, at the Quaker Square Hilton in Akron, Ohio. Sessions will deal with economics of the arts, historic preservation, and other cultural topics. Those interested in attending and/or presenting a paper should contact Dr. William S. Hendon, Association for Cultural Economics, Department of Urban Studies, University of Akron, Akron, OH 44325.

An exhibition of art created by members of the economics profession is planned. Economists are encouraged to submit their works of fiction, poetry, drama, musical scores, photography, oils, watercolors, etc. Those desiring to exhibit should send in advance copies, slides, or whatever is appropriate. While adequate security will be provided, exhibitors are expected to assume insurance and shipping costs. Please forward materials or requests for additional information to Dr. Douglas Shaw, Department of Urban Studies, University of Akron, Akron, OH 44325.

The International Health Economics and Management Institute has been founded as a nonprofit professional association to improve and expand knowledge in health economics and management systems worldwide. The Institute will provide a mechanism to gain better understanding of economic and health care problems, and to increase the opportunity for exchange of new ideas at both the theoretical and applied levels. Carolyne K. Davis, Administrator of the Health Care Financing Administration, is the Honorary Chairperson of an International Steering Committee, which is made up of representatives of six countries. For further information on the Institute and membership, contact John M. Virgo, IHEMI, Box 101, Southern Illinois University, Edwardsville, IL 62026-1001.

A "Conference on New Approaches to Modeling, Specification Selection, and Econometric Inference" will be held May 10–11 at The University of Texas-Austin. Those interested in participating should contact Dr. Raymond Smilor, Kozmetsky Endowment, 2815 San Gabriel Avenue, Austin, TX 78705.

The eleventh annual Conference of the European Association for Research in Industrial Economics (EARIE) will be held August 29–31, 1984, at the European Institute of Business Administration, Fontainebleau, France. Sessions include the economics of industry and firm organization, industrial and competition policy, industrial organization and international trade, industry studies, regulation of industries and firms, industrial organization theory, as well as other areas of current research interest in industrial economics. Those interested in presenting a paper must send a one-page abstract no later than March 15; complete papers must be received by May 15. Contact the Chairman of the Programme Committee: Dr. H. Landis, Gabel, INSEAD, Boulevard de Constance, 77305 Fontainebleau Cedex, France.

The fifth International Conference on Futures Analysis, Forecasting and Planning for Telecommunications will be held at the University of British Columbia Conference Centre, July 1–4, 1984, in Vancouver. For more information, contact J. Alleman, GTE Service Corporations, One Stamford Forum, Stamford, CT 06904 (telephone 203 + 965-2391), or D. S. Furman, Bell

Laboratories, Rm 5C-124, 600 Mountain Avenue, Murray Hill, NJ 07974 (telephone 201+582-2156).

The College of Business and Public Administration, University of Arizona, has established the Karl Eller Center for the Study of the Private Market Economy. The Center will serve three primary objectives: to support fundamental research on the private market economy; to develop and maintain an entrepreneurial studies program; and to promote meaningful interaction between the academic and corporate communities. A founding grant from Arizona businessman Karl Eller and additional funding from private sources will support Center activities. Address inquiries to Dean Kenneth R. Smith, College of Business and Public Administration, University of Arizona, Tucson, AZ 85721.

Economists who are strongly oriented toward the humanities, who use humanistic methods in their research, and who will be participating in meetings held outside the United States, Mexico, and Canada that are concerned with the humanistic aspects of their discipline are eligible to apply for small travel grants of the American Council of Learned Societies. Financial assistance is limited to air fare between major commercial airports and will not exceed one-half of projected economy-class fare. Social scientists and legal scholars who specialize in the history or philosophy of their disciplines are eligible if the meeting they wish to attend is so oriented. Applicants must hold a Ph.D. degree or its equivalent, and must be citizens or permanent residents of the United States. To be eligible, proposed meetings must be broadly international in sponsorship or participation, or both. The deadlines for applications to be received in the ACLS office are: meetings scheduled between July and October, March 1; for meetings scheduled between November and February, July 1; for meetings scheduled between March and June, November 1. Please request application forms by writing directly to the ACLS (Attention: Travel Grant Program), 800 Third Avenue, New York, NY 10022, setting forth the name, dates, place, and sponsorship of the meeting, as well as a brief statement describing the nature of your proposed role in the meeting.

Deaths

William J. Fellner, Resident Scholar, American Enterprise Institute, September 15, 1983.

Sidney Weintraub, professor of economics, University of Pennsylvania, June 18, 1983.

Retirements

Blaine E. Grimes, professor of economics, Ohio Wesleyan University, June 1983.

Wilfred Malenbaum, professor of economics, University of Pennsylvania, June 30, 1983.

Foreign Scholars

Shlomo Maital, Technion Institute of Technology, Haifa: visiting professor of applied economics, Sloan School of Management, Massachusetts Institute of Technology, Spring 1984.

Promotions

Roy D. Adams: professor of economics, Iowa State University, August 21, 1983.

Christophe P. Chamley: associate professor, department of economics, Yale University, July 1, 1983.

Patrick J. Corcoran: research officer and senior economist, Research and Statistics Function, Federal Reserve Bank of New York, August 1, 1983.

William P. Curington: associate professor of economics, University of Arkansas, July 1, 1983.

Charles Dale: research economist, U.S. Army Research Institute, Alexandria, VA, July 24, 1983.

Edward J. Frydl: assistant vice president, Research and Statistics Function, Federal Reserve Bank of New York, August 1, 1983.

William J. Gasser: assistant vice president, Research and Statistics Function, Federal Reserve Bank of New York, August 1, 1983.

John Geanakoplos: associate professor, department of economics, Yale University, July 1, 1983.

Roger Ginder: associate professor of economics, Iowa State University, July 1, 1983.

Robert J. Gitter: associate professor of economics, Ohio Wesleyan University, Fall 1983.

E. William Johnson: professor of economics, Shepherd College, August 16, 1983.

Randall H. King: associate professor of Economics, University of Akron, September 1983.

Nancy P. Marion: associate professor of economics, Dartmouth College, July 1, 1983.

John M. Marshall: professor of economics, University of California-Santa Barbara, July 1, 1983.

Joseph S. Mascia: associate professor, department of banking, economics and finance, Adelphi University, July 1983.

J. Peter Mattila: professor of economics, Iowa State University, July 1, 1983.

William M. Petersen: chief, Banking Studies Division, Banking Studies Department, Federal Reserve Bank of New York, July 28, 1983.

Karen L. Pickerill: assistant professor of economics, Western Kentucky University, August 16, 1983.

Seyed M. Raji: professor of economics, Dowling College, September 1, 1983.

James Schmidt: associate professor of economics, University of Nebraska-Lincoln, August 22, 1983.

John T. Scott: associate professor of economics, Dartmouth College, July 1, 1983.

Richard W. Stratton: associate professor of economics, University of Akron, September 1983.

Administrative Appointments

Syed Ahmad: chairman, department of economics, McMaster University, July 1983.

Milton Esbitt: chairperson, department of economics, business and political science, Rosary College, July 1983.

Carlos F. Hidalgo: Dean, Division of Business and Economics, Inter American University, Puerto Rico, fall 1983.

U Jin Jhun: chairman, department of economics, State University of New York-Oswego, September 1, 1983.

Robert R. Keller: chair, department of economics, Colorado State University, July 1, 1983.

Randall H. King: chairman, department of economics, University of Akron, June 1983.

Prem S. Laumas: chairman, department of economics, Northern Illinois University, July 1, 1983.

Paul W. MacAvoy: Dean, Graduate School of Management, University of Rochester, July 1, 1983.

Stephen P. Mullin: Budget Director, Budget Division, City of St. Louis, April 25, 1983.

Roger Riefler: chair, department of economics, University of Nebraska-Lincoln, August 15, 1983.

Paul Wachtel: chairman, department of economics, New York University, September 1983.

Samuel H. Williamson: chair, department of economics, Miami University (Ohio), August 1983.

Edward E. Zajac, Bell Laboratories: head, department of economics, University of Arizona, January 1, 1984.

Appointments

Donald Alexander: visiting assistant professor, department of economics, College of William and Mary, September 1983.

Robert Z. Aliber, University of Chicago: visiting professor of economics, Williams College, fall 1983.

James Anton: assistant professor, State University of New York-Stony Brook, fall 1983.

Madelyn Antoncic: economist, Domestic Financial Markets Division, Financial Markets Department, Federal Reserve Bank of New York, August 31, 1983.

Robert Archibald, College of William and Mary: visiting associate professor of economics, University of Arizona, 1983–84.

Alan Auerbach, Harvard University: associate professor of economics, University of Pennsylvania, July 1, 1983.

Truman F. Bewley, Northwestern University: professor, department of economics, Yale University, July 1, 1983.

Alok Bhargava, London School of Economics: assistant professor of economics, University of Pennsylvania, July 1, 1983.

Mario I. Blejer, International Monetary Fund: associate professor of economics and international business, New York University, September 1983.

Steven R. Booth, Cornell University: instructor, department of economics, Ohio Wesleyan University, September 1, 1983.

David G. Brown: professor of economics, Bellarmine College, August 29, 1983.

Paul Burkett: Syracuse University: visiting assistant professor of economics, University of Miami, August 1983.

Charles A. Calhoun: assistant professor of economics, Florida State University, August 5, 1983.

Nicholas Scott Cardell, Oak Ridge National Laboratory: associate economist, The Rand Corporation, August 1983.

Richard H. Clarida, Harvard University: assistant professor, department of economics, Yale University, July 1, 1983.

Judith Cox, University of Washington: visiting associate professor of economics, Williams College, 1983–84.

Keith W. Crane, Indiana University: associate economist, The Rand Corporation, August 1983.

William S. Custer, University of Illinois-Urbana: assistant professor of economics, Northern Illinois University, August 1983.

Michael L. Davis: visiting assistant professor of economics, Southern Methodist University, August 1983.

Anil Deolalikar, Yale University: assistant professor of economics, University of Pennsylvania, July 1, 1983.

Kathleen Donahoo: economist, Banking Studies Division, Banking Studies Department, Federal Reserve Bank of New York, August 22, 1983.

Kenneth G. Elzinga, University of Virginia: visiting professor of economics, Trinity University, spring 1984.

David W. Emmons, Virginia Polytechnic Institute and State University: visiting assistant professor, Wayne State University, August 30, 1983.

Arlene Eskin: economist, Developing Economies Division, External Financing Department, Federal Reserve Bank of New York, August 3, 1983.

Arturo Estrella: economist, Domestic Financial Markets Division, Financial Markets Department, Federal Reserve Bank of New York, August 1, 1983.

Roger Farmer, University of Western Ontario: assistant professor of economics, University of Pennsylvania, July 1, 1983.

Stefano Fenoaltea: visiting associate professor of economics, Williams College, spring 1984.

Harold O. Fried, Tulane University: visiting assistant professor of economics, Union College, September 1983.

Joyce Gleason: assistant director, Center for Economic Education, University of Nebraska-Lincoln, August 22, 1983.

Harold T. Gross, University of Texas-Dallas: research scientist, John Gray Institute, Lamar University, September 1, 1983.

J. Arne Hallam: assistant professor, department of economics, Iowa State University, July 1, 1983.

Robert O. Heavner, United States Air Force: economist, The Rand Corporation, July 1983.

Jack C. High, George Mason University: California State University-Long Beach, August 29, 1983–May 31, 1984.

Gregory G. Hildebrandt, United States Air Force: economist, The Rand Corporation, Washington D.C. and Santa Monica, September 1983.

Robert Holland, Purdue University: instructor, Miami University (Ohio), August 1983.

Frank T. Hollenbeck: instructor, department of economics, Western Kentucky University, fall 1983.

Bengt R. Holmstrom, Northwestern University: professor of economics and management, Yale University, July 1, 1983.

John H. Huston, University of Wisconsin-Madison: assistant professor of economics, Trinity University, August 1982.

Robert M. Jennings, Jr.: assistant professor of accounting, Bellarmine College, August 29, 1983.

David Kelsey: visiting assistant professor of economics, Southern Methodist University, August 1983.

Waseem A. Khan, University of Southern Colorado: international economist, First Interstate Bank, Los Angeles.

Benjamin Kim: assistant professor of economics, University of Nebraska-Lincoln, August 22, 1983.

H. Youn Kim: assistant professor of economics, Western Kentucky University, fall 1983.

Elizabeth King: associate economist, The Rand Corporation, August 1983.

Kenneth M. Kletzer, University of California-Berkeley: assistant professor, department of economics, Yale University, July 1, 1983.

Thomas F. Kompas: assistant professor, department of economics, Iowa State University, June 1, 1983.

Mark N. Kosmo, University of Wisconsin: visiting assistant professor of economics, Dartmouth College, August 1, 1983.

Daniel E. Laufenberg: visiting associate professor, department of economics, Iowa State University, August 21, 1983.

Miles C. Leahey, Data Resources, Inc.: senior economist, Lehman Brothers, September 6, 1983.

Thomas P. Lyons, Cornell University: assistant professor of economics, Dartmouth College, July 1, 1983.

Mark McBride: assistant professor, Miami University (Ohio), August 1983.

Jane E. Mather, Stanford University: instructor in economics, Dartmouth College, July 1, 1983.

Paul R. Milgrom, Northwestern University: professor of economics and management, Yale University, July 1, 1983.

Michael R. Myler, Albion College: assistant professor of economics, Mount Union College, August 1, 1983.

Julianne Nelson, Princeton University: instructor, Graduate School of Business Administration, New York University, September 1983.

Alfonse Novales: assistant professor, State University of New York-Stony Brook, fall 1983.

Prasanta K. Pattanaik: professor of economics, Southern Methodist University, August 1983.

Thomas S. Paynter, University of California-Berkeley: assistant professor of economics, Northern Illinois University, August 1983.

David G. Pearce, Princeton: assistant professor, department of economics, Yale University, July 1, 1983.

Ronnie J. Phillips, Texas A&M University: assistant professor of economics, Colorado State University, August 20, 1983.

Susan Porter-Hudak, University of Wisconsin-Madison: assistant professor of economics, Northern Illinois University, August 1982.

Kathleen L. Reid, Pennsylvania State University: instructor in economics, Shepherd College, August 16, 1983.

Vincent R. Reinhart: economist, Industrial Economies Division, International Research Department, Federal Reserve Bank of New York, July 13, 1983.

John R. Schroeter: assistant professor, department of economics, Iowa State University, August 21, 1983.

Clinton R. Shiells: assistant professor of economics, California State University-Long Beach, August 29, 1983–May 30, 1985.

Martha E. Shiells: assistant professor of economics, California State University-Long Beach, August 29, 1983–May 30, 1985.

Joaquim Silvestre, Universitat Autonoma de Barcelona: professor, department of economics, University of California-Davis, July 1, 1983.

Allen Sinai, Data Resources, Inc.: senior vice president and chief economist, Lehman Brothers, September 6, 1983.

Davinder Singh: associate professor of economics, California State University-Long Beach, August 29, 1983.

Padmanabhan Srinagesh, University of Illinois-Chicago: visiting assistant professor of economics, Williams College, 1983–84.

Dana Stevens, New College, University of Southern Florida: visiting associate professor of economics, Williams College, fall 1983.

Joseph S. Tracy, University of Chicago: assistant professor, department of economics, Yale University, July 1, 1983.

Robert G. Vice: assistant professor of economics, California State University-Long Beach, August 29, 1983–May 30, 1984.

Xavier Vives, University of California-Berkeley: assistant professor of economics, University of Pennsylvania, July 1, 1983.

Robert C. Vogel, Syracuse University: professor of economics, University of Miami, August 1983.

Bernard L. Weinstein, University of Texas-Dallas: assistant director for research and policy, John Gray Institute, Lamar University, June 1, 1983.

Yoram Weiss: professor, State University of New York-Stony Brook, fall 1984.

Gary W. Williams: assistant professor, department of economics, Iowa State University, July 1, 1983.

Oliver E. Williamson, University of Pennsylvania: professor of law, management and economics, Yale University, July 1, 1983.

John D. Worrall: associate professor of economics, Rutgers University, July 1, 1983.

Edward E. Zajac, Bell Laboratories: professor of economics, University of Arizona, January 1, 1984.

Thomas W. Zuehlke: assistant professor of economics, Florida State University, August 5, 1983.

Leaves for Special Appointments

Craig MacPhee, University of Nebraska-Lincoln: visiting international economist, U.S. Department of Labor, August 23, 1983–August 21, 1984.

Wallace Peterson, University of Nebraska-Lincoln: visiting professor, University of Nevada-Las Vegas, August 22, 1983–May 12, 1984.

Resignations

John J. Beggs, Yale University: Australian National University, July 1, 1983.

Thomas R. Dalton, Northern Illinois University: Berry College, August 1983.

Albert Fishlow, Yale University: University of California-Berkeley, July 1, 1983.

William R. Keeton, Yale University: Federal Reserve Bank of Kansas City, July 1, 1983.

Nicholas R. Lardy, Yale University: University of Washington, July 1, 1983.

Richard J. Murnane, Yale University: Graduate School of Education, Harvard University, July 1, 1983.

John C. Soper, Northern Illinois University: John Carroll University, August 1982.

Kenneth I. Wolpin, Yale University: Ohio State University, July 1, 1983.

NOTE TO DEPARTMENTAL SECRETARIES AND EXECUTIVE OFFICERS

When sending information to the *Review* for inclusion in the Notes Section, please use the following style:

A. Please use the following categories:

1—Deaths
2—Retirements
3—Foreign Scholars (visiting the USA or Canada)
4—Promotions
5—Administrative Appointments
6—New Appointments
7—Leaves for Special Appointments (NOT Sabbaticals)
8—Resignations
9—Miscellaneous

B. Please give the name of the individual (SMITH, Jane W.), her present place of employment or enrollment: her new title (if any), new institution and the date at which the change will occur.

C. Type each item on a separate 3×5 card and please do not send public relations releases.

D. The closing dates for each issue are as follows: *March*, October 15; *June*, January 15; *September*, April 15; *December*, July 15.

All items and information should be sent to the Assistant Production Editor, *American Economic Review*, Room 8279, Bunche Hall, University of California, Los Angeles, CA 90024.